Check for updates

American Educational Research Journal
April 2013, Vol. 50, No. 2, pp. 397–428
DOI: 10.3102/0002831212469045
© 2012 AERA. http://aerj.aera.net

Evaluating Math Recovery: Assessing the Causal Impact of a Diagnostic Tutoring Program on Student Achievement

Thomas M. Smith
Paul Cobb
Dale C. Farran
David S. Cordray
Vanderbilt University
Charles Munter
University of Pittsburgh

Mathematics Recovery (MR) is designed to identify first graders who are struggling in mathematics and provide them with intensive one-to-one tutoring. We report findings from a 2-year evaluation of MR conducted in 20 elementary schools across five districts in two states. The design allowed for the estimation of the counterfactual growth trajectory based on those students

---

Thomas M. Smith is an associate professor of public policy and education in the Department of Leadership, Policy, and Organizations, Peabody College, Vanderbilt University, #414 GPC 230 Appleton Place, Nashville, TN 37203-5721, USA; e-mail: *thomas.smith@vanderbilt.edu*, and Director of the National Center on Scaling Up Effective Schools. His research focuses on scaling up effective practices and on improving teaching and learning at scale.

Paul Cobb is a professor of mathematics education in the Department of Teaching and Learning, Peabody College, Vanderbilt University. His current research focuses on improving mathematics teaching and learning at scale.

Dale C. Farran is a professor in the Department of Teaching and Learning, Peabody College, Vanderbilt University, and senior associate director of the Peabody Research Institute. Her research focuses on early intervention for children at risk for school failure due to poverty and/or disabilities.

David S. Cordray is a professor of public policy and professor of psychology in the Department of Psychology and Human Development, Peabody College, Vanderbilt University. His research focuses on the development of methodological refinements of randomized field trials, quasi-experimental designs, meta-analysis, and nontraditional forms of casual inquiry.

Charles Munter is an assistant professor of mathematics education in the School of Education at the University of Pittsburgh. His research focuses on specifying high-quality forms of mathematics instruction and supporting teachers' development of such practices.

randomly assigned either to a tutoring cohort with a delayed start or to a wait list. Results demonstrate strong end of first grade effects on a diagnostic measure developed by MR and weak to moderate effects (effect size, .15–.30) on measures administered by external evaluators. By the end of second grade, no significant effects were found on any measures. Practical and research implications are discussed.

KEYWORDS: elementary mathematics, experiments, tutoring, diagnostic assessment, intervention

Children enter school with a wide range of mathematical abilities (Baroody, 1987; Griffin & Case, 1999; Houssart, 2001; Wright, 1991, 1994). These initial gaps in student achievement persist, and in many cases widen, as students progress through the elementary grades and into middle school (Aunola, Leskinen, Lerkkanen, & Nurmi, 2004; Cockcroft, 1982; Duncan, Claessens, & Engel, 2004; Princiotta, Flanagan, & Germino Hausken, 2006; Stevenson & Newman, 1986). Students from traditionally low-performing subgroups (low socioeconomic status, Black or Hispanic, English language learners, and students with disabilities) are at particular risk of starting school behind in mathematics (Alexander & Entwistle, 1998; Barton, 2003; Berends, Lucas, Sullivan, & Briggs, 2005; Crosnoe, 2005; Wilms, 1986). Further, research on tracking suggests that students who fall behind in student achievement tend to have less access to the kinds of learning opportunities provided to students at the middle or higher ends of the achievement distribution (Oakes, 1990; Schmidt et al., 1999/2002). The consistent finding that differences in early mathematical abilities are relatively stable and can lead to students being tracked into classes of differing instructional quality in later elementary and middle grades indicates that there is pressing need to equip schools with effective methods for closing the gap early in children's school careers (Aubrey, Dahl, & Godfrey, 2006; McWayne, Fantuzzo, & McDermott, 2004).

Among 20 programs listed as early childhood and elementary mathematics interventions (including curricula) in the What Works Clearinghouse (WWC), only 2—Building Blocks for Math (Clements & Sarama, 2007) and Pre-K Mathematics (Starkey & Klein, 2005)—show "positive effects" for improving student learning (http://ies.ed.gov/ncee/wwc/). Both programs are classroom-level interventions designed to help all students, including those at risk of school failure, solve problems presented verbally and without objects by representing and abstracting from their informal experiences with quantities. Here we report on an evaluation of Mathematics Recovery (MR), an intervention program that aims to reduce the achievement gap in mathematics by providing low-performing first graders with intensive one-on-one tutoring with a certified teacher who has been trained to continuously assess students' current developmental level and engage them in tasks to advance them in each of six aspects of arithmetical knowledge.

MR shares a number of features with Building Blocks and Pre-K Mathematics, including the use of research-based developmental progressions to design curricula and professional development. A developmental progression delineates a series of levels that children typically pass through in developing a particular skill or concept (e.g., reciting standard number word sequences, place value). Like Building Blocks and Pre-K Mathematics, MR provides developmentally sequenced instructional activities, assessment benchmarks, and professional development designed to enable teachers to use instructional task to move children through successive developmental levels. Clements and Sarama (2012) recently observed that the use of developmental progressions "in designing curricula and professional development may be responsible for the success of many early mathematics projects" (p. 969). More generally, current research in mathematics education emphasizes the design and implementation of instructional sequences grounded in research-based developmental progressions at all grade levels and at the college level (Mosher, 2011). Research on the development of students' thinking in different mathematical domains also informed the development of the Common Core State Standards for Mathematics (Common Core State Standards Initiative, 2010).

MR is unique among the 34 early mathematics interventions (including Building Blocks and Pre-K Mathematics) reviewed by Wagner (2005) in featuring individualized one-on-one instruction. Although MR was first implemented in the United States in 1995 and has served approximately 6,000 students in 25 states, its effectiveness has not been assessed using experimental methods and externally validated student outcome measures. There are several reasons to suggest that MR is likely to have a positive impact on closing early mathematics achievement gaps.

First, it is well documented that well-targeted and well-designed one-to-one tutoring can be effective in supporting students' mathematical learning (for a meta-analysis, see Cohen, Kulik, & Kulik, 1982). Furthermore, prior research suggests greater improvement in achievement when tutors are certified teachers rather than paraprofessionals, volunteers, or peers (for a review, see Wasik & Slavin, 1993).

Second, MR is consistent with several characteristics of effective tutoring identified in the literature, including use of concrete materials to promote conceptual learning (Fuchs et al., 2005), use of appropriate prompting (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001), allowing students to reach an impasse (VanLehn, Siler, Murray, Yamauchi, & Baggett, 2003), appropriately introducing tasks to the students (McArthur, Stasz, & Zmuidzinas, 1990), and developing explicit rather than merely intuitive models of student reasoning (Derry & Potts, 1998; Lesh & Kelly, 1997).

Third, MR involves diagnostic tutoring in which tutors adapt instruction to the child's current level of competence. In this regard, MR differs from Number Rockets, a small-group tutoring intervention for Grade 1 students at risk for mathematics difficulties (Fuchs et al., 2005). Number Rockets

tutors use scripts to maintain consistency as adult tutors work with groups of students at risk for learning disabilities. In contrast, MR tutors are trained to use both Learning and Instructional Frameworks (detailed in the following) to select tasks appropriate for a student's current level of mathematical reasoning, and MR conducts tutoring in one-to-one format, not groups. Evidence from an effectiveness trial of Number Rockets, implemented among at-risk first-grade students in 76 schools in four urban districts, showed that participants had significantly higher Test of Early Mathematics Ability (TEMA-3) scores than at-risk Grade 1 students in the control group (effect size = 0.34, $p < .001$; Rolfhus et al., 2012). A priori, we conjectured that MR tutors' approach of individualizing instruction based on ongoing assessments of the student's reasoning might be more effective in supporting the learning of low-performing students. The demonstrated effectiveness of Building Blocks, while not a tutoring program, provides the most direct evidence of the value of early mathematics instruction that is informed by knowledge of children's mathematical reasoning. Clements and Sarama (2012), the developers of Building Blocks, concluded that developmentally appropriate teaching requires knowledge of "the mathematical content (goal), the developmental progressions of children's thinking and learning, and instructional tasks and teaching strategies that help children move along those progressions" (p. 970). Training provided to MR tutors emphasizes each of these three aspects of teacher knowledge.

Finally, MR's grounding in cognitive research on early number learning is a primary reason for anticipating that the program has the potential to produce increases in the mathematics achievement of the lowest achieving first graders. The MR program comprises an elaborated body of research-based developmental progressions, instructional practices for supporting the development of children's numerical reasoning, and a set of professional development activities for supporting tutors' acquisition of this knowledge and development of these practices. The developers of MR have codified this knowledge in a Learning Framework and an Instructional Framework that tutors use to assess children's arithmetical knowledge and to tailor instruction to their current levels of arithmetical reasoning (Phillips, Leonard, Horton, Wright, & Stafford, 2003; Wright, 2003; Wright, Martland, & Stafford, 2000; Wright, Martland, Stafford, & Stanger, 2002). The six aspects of early number knowledge included in the Mathematics Recovery Learning Framework are shown in Table 1.

Robert Wright, the lead developer of MR, formulated the initial version of the Learning Framework in 1990 (Wright, 1991, 1994). Wright and his colleagues delineated developmental progressions for four early numerical skills and concepts by drawing on cognitive models of children's numerical reasoning proposed by a number of researchers (Baroody, 1987; Baroody & Ginsburg, 1986; Carpenter & Moser, 1982, 1984; Clements, 1999; Fuson, 1988, 1992; Steffe, Cobb, & von Glasersfeld, 1988; Steffe, von Glasersfeld,

Table 1
Aspects of Early Number Knowledge Included in Math Recovery (MR) Learning Framework

Aspect of Early Number Knowledge	Definition	Stages/Levels
Stage of Early Arithmetical Learning (SEAL)	Strategies for solving early number tasks, differentiated by degrees of mathematical sophistication	Stage 0, at which a student cannot count visible items, to Stage 5, at which a student is able to use a range of non-count-by-one strategies
Forward number word sequence (FNWS)	The range of number words a student can correctly voice in increasing sequence	Level 0, at which a student cannot produce the FNWS from "one" to "ten," to level 5, at which a student is facile in FNWS up to "one hundred"
Backward number word sequence (BNWS)	The range of number words a student can correctly voice in decreasing sequence	Level 0, at which a student cannot produce the BNWS from "ten" to "one," to level 5, at which a student is facile in BNWS up to "one hundred"
Numeral identification	The range of numeric symbols a student can recognize and name	Level 0, at which a student cannot identify some or all of the numerals from "1" to "10," to level 4, at which a student can identify all numerals up to "1000"
Base ten arithmetical strategies	The extent to which a student can count by ones *and* tens	Level 1, at which a student has no concept of 10 as a unit, to level 3, at which a student can solve tasks by adding or subtracting units of tens and ones.
Structuring number	The extent to which a student can employ combining and partitioning strategies	Level 0, at which a student can subitize quantities up to only 3, to level 5, at which a student is able to utilize various number structures in a range of 1 to 20 without counting.

Richards, & Cobb, 1983). The Learning Framework specifies up to six developmental levels for each of six aspects of arithmetical knowledge shown in Table 1. The levels for each strand span children's typical development in arithmetic during the first 2 to 3 years of school. The Learning Framework is designed to assist tutors in diagnosing students' current understandings and strategies by providing information about students' responses in terms of levels of sophistication. The tutor uses this framework to create (and update after each tutoring session) a profile for each student's current arithmetical strategies and knowledge across the six aspects of early number knowledge.

The initial version of the Instructional Framework was developed in 1999. Wright and his colleagues developed instructional tasks appropriate for children as different levels of the Learning Framework drawing on research on early number instruction (Baroody, 1990; Beishuizen, 1993; Carpenter, Franke, Jacobs, Fennema, & Empson, 1997; Clements, 1999; Cobb, Gravemeijer, Yackel, McClain, & Whitenack, 1997; Fuson, 1990; Fuson, Wearne, et al., 1997; Hiebert & Wearne, 1992). The tutor uses the Instructional Framework to develop an instructional plan that is based on the child's profile. The Instructional Framework specifies at least six exemplary "teaching procedures" for each of 30 key topics in arithmetic that map onto the levels of the six aspects of early number knowledge in the Learning Framework. Each teaching procedure consists of a set of tasks accompanied by a description of the appropriate tools or manipulatives to use when posing the tasks. Additionally, each procedure provides detailed guidance for the tutor, including a statement of its purpose, suggestions for posing tasks, and a detailed discussion of possible student responses.

MR is the only early arithmetic program implemented in the United States that involves diagnostic, one-to-one tutoring (Wagner, 2005), a resource-intensive but potentially effective means of raising achievement (Cohen et al., 1982). Further, the majority of programs positively reviewed by the WWC begin at pre-K, while MR is designed to help close the math gaps that have already developed for many children by the start of first grade. The reports of MR tutors have consistently been positive. To date, the few studies that have attempted to directly measure its effects have used in-house measures that have not been externally validated (MacLean, 2003; Phillips et al., 2003). In this article, we report the results of a 2-year multisite randomized field trial to assess the effectiveness of the MR program on improving first-grade students' mathematics achievement. The specific research questions we addressed were:

Research Question 1: Did participation in MR raise the mathematics achievement of first-grade students identified as eligible for MR?

Research Question 2: If so, did participating students maintain the gains made in first grade through the end of second grade?

Design of the MR Program

The MR program consists of three components: (1) tutor training, (2) student identification and initial assessment, and (3) one-to-one tutoring. A logic model for the MR is depicted in Figure 1. The first component of the MR program, tutor training, involves 60 hours of professional development provided by an MR leader. The goal of this training is to support tutors in understanding the Learning and Instructional Frameworks and in developing new practices for clinical assessment and intervention teaching that involve adjusting instruction based on cognitive evaluations of the student's responses. The training typically involves an initial 5-day workshop during the summer followed by a 3-day workshop conducted approximately 3 weeks into the school year after trainees have completed assessment interviews with students selected for the program. The trainees in a district are also expected to meet as a cohort for 2 hours each month. The MR leader responsible for their training typically attends three of these meetings and also conducts individual coaching sessions with the MR trainees during these site visits. An overriding goal of the training is to teach tutors to use the MR Learning and Instructional Frameworks to assess children's arithmetical knowledge and to adjust instruction based on these evaluations by selecting tasks that are tailored to their current levels of arithmetical reasoning.

In the second component of the program, student identification and initial assessment, the tutor identifies students in his or her school who are potential candidates for MR by conducting a 10-minute initial screening interview with all first graders who have been identified by classroom teachers as struggling in mathematics. The initial screening interview consists of eight tasks involving numbers in the range 1 to 10 (although other instruments can be used, particularly if a district already employs an assessment for a similar purpose). The tasks focus on counting visible objects, forward and backward number word sequences, numeral identification, numeral sequencing, and elementary addition and subtraction. The pool of students identified as potential candidates for MR varies depending on the needs of the school, but the developers report that it is typically the lowest 20% to 25% of the first-grade population. The tutor then conducts a more extensive video-recorded MR 1.1 Assessment interview with each child in the pool. The results of these interviews are used to select children for one-on-one tutoring.

The third component of the program, one-to-one tutoring, is diagnostic in nature and focuses instruction at the current limits of each child's arithmetical reasoning (Baroody & Tiilikainen, 2003; Keogh, Major, Reid, Gandara, & Omori, 1978; Tang & Ginsburg, 1999). Tutors construct a profile of each selected child's current arithmetical knowledge based on the MR 1.1 Initial Assessment prior to conducting 4 to 5 30-minute one-to-one tutoring sessions with each child each week for approximately 12 weeks. Tutors work

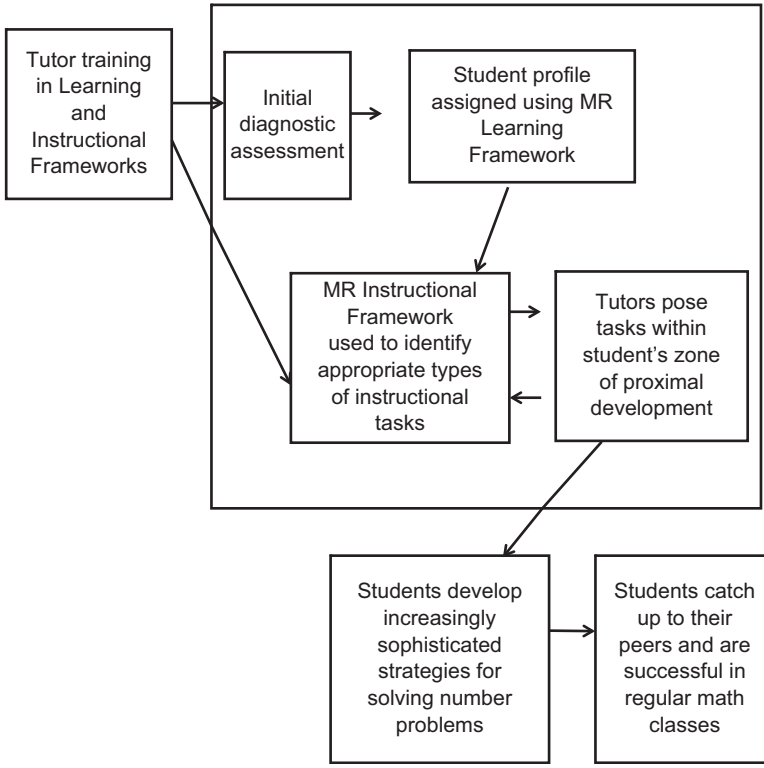


Figure 1. **Math Recovery logic model.**
Note. MR = Math Recovery.

with 3 or 4 students in each 12-week tutoring cycle, for a total of between 9 and 12 students each year. A tutor's selection of instructional tasks for a particular student is initially informed by the MR 1.1 Initial Assessment interview and later by ongoing assessments of the student's responses to instructional tasks posed in tutoring sessions. Tutors video-recorded all sessions for later review, during which they analyze the child's responses using the Learning Framework and then select appropriate types of tasks for the next session by using the Instructional Framework. The goal is to choose tasks that will be challenging but that the child has a reasonable chance of solving, thus avoiding undue frustration. To support engagement, tutors are supposed to give the child sufficient time to complete a task.

MR's training program for tutors lasts 1 year and costs districts $5,500 per teacher ($4,840 if two teachers or more from a single district receive training). The implementation of MR tutoring is flexible and can be tailored to

fit a district's needs. In the most common implementation, one teacher per school is trained as an MR specialist and receives a 50% release from classroom teaching to tutor the school's lowest performing first graders. Tutors typically work with 9 students during their training year and 12 students per year thereafter. The cost of the 50% classroom release, typically $25,000 to $30,000 per year, must be considered when weighing the cost against MR's potential benefits. Assuming that one tutor in a district receives MR training and works with 9 students during the training year and then 12 students in each of the following 2 years, the cost per student over this 3-year period is $2,500 to $3,000 per student, slightly more than the $2,000 per student allocated through the supplemental education services (SES) provision of the No Child Left Behind Act of 2001 (Burch, 2007).

Research Design

Study Population

The 2-year evaluation of Mathematics Recovery was conducted in 20 elementary schools (5 urban, 10 suburban, and 5 rural) across five districts in two states. Each school was a "fresh site" in that MR was implemented for the first time for the purposes of the study. The participating districts recruited 18 teachers to receive training and participate as MR tutors. All had at least 2 years of classroom teaching experience. Of the tutors, 16 received half-time teaching releases to serve one school each; 2 of the tutors in the urban district received full-time teaching releases to serve two schools each. All 18 tutors received MR training during the first year of the study at one of two sites, and all remained in the study for the full 2 years. Our original design was to have tutors work with 9 students per school (3 per 12-week cycle) in the first year and 12 per school (4 per 12-week cycle) in the second year. Because of local demands on tutors' time, however, 17 out of 18 tutors worked with 9 students per school in both years. One tutor worked with 6 students in Year 2 because of a maternity leave.

Tutors selected students for participation at the start of first grade based on their performance on the initial screening interview and the follow-up MR 1.1 Assessment interview. The number of students deemed eligible for tutoring ranged from 17 to 36 per school. The number of study participants before attrition totaled 517 in Year 1 and 510 in Year 2, of whom 172 received tutoring in Year 1 and 171 received tutoring in Year 2 (the methods for determining which students received tutoring in each tutoring cycle are described in the following section).[1] The Year 1 participants were followed until the end of second grade, and the Year 2 participants were followed until the end of first grade. As shown in Table 2, approximately 50% of Year 1 participants were males, 46% were non-White, 17% were classified as limited English proficient (LEP), and 72% received free or reduced lunch (FRPL).

Table 2
Sample Characteristics of Math Recovery–Selected First Graders by Cohort

	WJ Math Reasoning (SS)	Percentage Female	Percentage Minority	Percentage LEP	Percentage FRPL
Year 1 (Fall 2007)					
All	92.8	51.1	46.3	16.8	71.8
Treatment	93.2	48.3	46.2	16.3	72.5
Control	92.4	53.5	46.5	17.2	71.2
Year 2 (Fall 2008)					
All	94.8	58.7	49.9	12.3	57.5
Treatment	95.7	61.4	46.5	8.8	55.6
Control	94.3	56.9	52.2	14.6	58.9

Note. SS = scale score; LEP = limited English proficient; FRPL = percentage of students receiving free or reduced price lunches.

For reasons we have not been able to determine, the Year 2 study participants tended to be more economically advantaged compared to Year 1 study participants (58% received free or reduced price lunches compared to 72% in Year 1) and were less likely to be classified as limited English proficient (12% compared to 17%), even though tutors drew participants from the same sample schools in both years. Across both years, however, there were no statistically significant differences between the treatment and control groups on gender, minority status, LEP, FRPL, or prior achievement.

Study Design

The structure of the MR program allowed us to use the fact that two thirds of the participating students had their treatment delayed by either 12 or 24 weeks to establish an experimentally assigned control group for each cohort of participants that consisted of both students whose treatment had not yet begun and the students who were assigned to a "wait list." We established the essential characteristics of an experimental design at the beginning of each year by randomly assigning the students identified as eligible for MR tutoring to one of the three treatment cohorts with different start dates (i.e., Cohort A–September, Cohort B–December, or Cohort C–March) or to the wait list. We also ensured that students in treatment cohorts were not pulled for tutoring during mathematics classes. We could therefore compare change in students' mathematics achievement during their 12 weeks of participation in MR to the gains they would have made if they had not participated in the intervention. In both years, students on the randomly

ordered waiting list were selected to join an MR tutoring cohort if a student assigned to a cohort left the school, was deemed "ineligible" due to a special education placement, or received another math-focused service. Provided the attrition was unrelated to participation in MR, replacement of a student originally assigned to a tutoring cohort with one on the waitlist maintains the treatment-control balance constructed by the original random assignment. This is likely to be the case because the most common reason for attrition (described in greater detail in the following) was withdrawal from a study school, an action unlikely to be related to their participation (or nonparticipation) in MR.

In reporting our primary results, we use the students who remained on the waitlist (i.e., those who remained in the study and were not moved into a cohort) at the end of first grade as a control group when estimating the effects of participating in MR on both end of first grade and end of second grade achievement. We then make use of the delay in treatment across cohorts to estimate the impact of MR participation in Cohorts A, B, or C on achievement growth trajectories from the beginning of first grade to the end of second grade. These latter analyses highlight the relative impact of MR compared to the overall growth in achievement that nonparticipants made over five time points between the start of first grade and the end of second grade.

Fidelity of Implementation

In order to fully understand the effect of the intervention, we conducted an assessment of implementation fidelity.[2] As described in greater detail by Munter (2010), our assessment of implementation fidelity focused on the extent to which tutors' practices as they implemented the program were consistent with those intended by the developers of the MR program (Dusenbury, Brannigan, Falco, & Hansen, 2003). We developed a coding scheme to assess tutors' adherence to *structural* features of the program (e.g., Was the initial assessment protocol followed correctly? Were tutoring sessions of the expected length of time? Was an appropriate proportion of tutoring time devoted to each of the six aspects of arithmetical knowledge?). In addition, we consulted with the program developers to operationalize the key *process* aspects of MR tutoring, including (a) the ongoing assessment of the child's thinking and strategies and (b) the selection of tasks based on these assessments. We then developed coding schemes for assessing these process aspects of MR tutoring (e.g., Did the tutor pose sets of related tasks that were in accord with those specified in the MR Frameworks? Did the tutor implement the tasks as intended by program developers?)

We hired five coders who received a 5-day version of MR training from the same MR personnel who trained the study tutors, followed by training on the fidelity coding scheme. We then worked with the coders to achieve an

observer-agreement percentage of at least 80% on all classes of codes combined. Once reliability was achieved, the coders coded video-recordings of the initial assessments and tutoring sessions conducted with 107 (32%) of the 343 students who received tutoring. For each randomly selected student, coders assessed the fidelity with which the initial assessment and 12 tutoring sessions (drawn from across an entire tutoring cycle) were conducted. For ongoing reliability checks, approximately 20% of the videos analyzed in the fidelity study were blindly double-coded, after which discrepancies were resolved. (The reader is referred to Munter, 2010, for a detailed description of our methods for assessing implementation fidelity).

Outcome Measures

We used both proximal and distal assessments to measure the participating children's mathematical knowledge at different points during the study. A listing of the instruments and the testing schedule is shown in Table 3. MR tutors assessed the students participating in the study using the MR Initial Assessment (MR1.1) when they identified students who were eligible of tutoring.[3] The content of the MR1.1 closely parallels the content of MR tutoring tasks. In order to have an additional proximal assessment administered by trained external assessors, we worked with the program developers to design "MR proximal assessments," one appropriate for the end of first grade (MRp) and one appropriate for the end of second grade (MRp2). Both assessments contained 12 items. The first 2 tasks on the first-grade assessment focused on number word sequences ("What is the number after – ?") and numeral recognition (the assessor shows the child a numeral and asks, "What number is this?"). The remaining 10 tasks on the first-grade assessment and all 12 tasks on the second-grade assessment required arithmetical calculations (e.g., 6 + 2 = _; 78 – _ = 73) and were presented either as written numerical expressions that the assessor read to the child or as real-world word problems. The assessor read tasks presented in both formats to the child. We developed three alternative forms of the first-grade assessment by changing the numbers while preserving the type and format of tasks so we could give the assessment multiple times during the year without repeating exactly the same items.[4]

The trained external assessors also assessed each participating student on alternating forms of a more distal measure, the Woodcock Johnson III Achievement (WJIII) subtests Applied Problems (WJ AP), Quantitative Concepts (WJ QC), and Math Fluency (WJ MF). We chose these subtests of the WJ because they (a) closely map to the central aspects of arithmetical knowledge delineated in the research literature and specified in the MR Learning Framework, (b) are scaled over a long enough developmental period to use in two successive grades for children this age, (c) have alternating forms so we could give the tests multiple times without repeating

Table 3
Math Recovery (MR) Treatment (Tx) and Assessment Cycle for Each School

	Start First Grade	Tx A (Fall)	Assessment	Tx B (Winter)	Assessment	Tx C (Spring)	End First Grade	End Second Grade
Cohort A	MR1.1	T_x	MR1.1		MR1.1		MR1.1	MRp2
(n = 3)	MRp		MRp		MRp		MRp	WJ full
	WJ full		WJ full		WJ MF		WJ full	
Cohort B	MR1.1		MR1.1	T_x	MR1.1		MR1.1	MRp2
(n = 3)	MRp		MRp		MRp		MRp	WJ full
	WJ full		WJ full		WJ full		WJ full	
Cohort C	MR1.1		MR1.1		MR1.1	T_x	MR1.1	MRp2
(n = 3)	MRp		MRp		MRp		MRp	WJ full
	WJ full		WJ MF		WJ full		WJ full	
Wait list	MR1.1		MR1.1		MR1.1		MR1.1	MRp2
($7 \leqslant n$	MRp		MRp		MRp		MRp	WJ full
$\leqslant 16$)	WJ full		WJ MF		WJ MF		WJ full	

Note. n = the size of the sample in each cohort or on the wait list in each of the 20 participating schools; MR1.1 = MR Initial Assessment (given by MR tutor in school or an experienced MR tutor from a non-study school hired by the U.S. Math Recovery Council); MRp = Proximal Math Recovery Assessment (developed by study team in collaboration with MR developers); WJ full = Full Woodcock Johnson III Achievement Assessment–Applied Problems (WJ AP), Math Fluency (WJ MF), and Quantitative Concepts (WJ QC); WJ MF = Math Fluency subtest of the Woodcock Johnson III Achievement Assessment only.

exactly the same items, and (d) have demonstrated correlation with other validated tests. The median reliability coefficient alphas for all age groups for the standard battery of the WJIII tests range from .81 to .94. Applied Problems and Quantitative Concepts together constitute a measure of Mathematics Core Reasoning (WJ MR). Each of these two subtests takes 10 to 15 minutes to administer, while Fluency is a timed measure that takes 3 minutes.

We hired and trained local assessors who administered all assessment instruments except the MR1.1 to students one-on-one. They administered the full battery of all instruments to all students at the beginning and end of their first-grade year and at the end of second grade for students who participate in the first year of the study (see Table 3). In addition, the assessors administered the full battery to all students when they entered and exited tutoring. They administered a partial battery comprising the MRp and the WJ MF subtest to the other students participating in the study at these intermediate time points. This testing schedule allowed us to both minimize the testing burden of those students who were not entering or exiting tutoring while allowing us to estimate both the counterfactual growth trajectory over five time points between the beginning of first grade and the end of second grade. Thus, students on the wait list took the WJ AP and WJ QC subtests two times in first

grade, students in Cohorts 1 and 3 took them three times, and students in Cohort 2 took them four times. If a testing/practice effect confound exists, we would expect the Cohort 2 WJ AP and WJ QC scores to be greater than those for Cohorts 1 and 3 (holding constant respective Time 1 assessment score). While this was the case for WJ QC, the differences were not statistically significant (Cohort 2 vs. 1, p = .311; Cohort 2 vs. 3, p = .477). Although the Cohort 2 predicted gain for WJ AP exceeded that of cohort 1 (p = .041), the Cohort 2 predicted gain was less than that of Cohort 3 and was not statistically significant (p = .465). Thus, there is no strong evidence of a testing/practice effect.

We expected MR to have the largest impact on the MR1.1 because the tasks are similar to those used in tutoring sessions and are presented one-on-one by the tutor who has worked with the child using similar materials. We also expected MR to have a stronger impact on the MR proximal (MRp and MRp2) assessments (particularly in first grade) than on the WJIII subtests because they were specifically designed to assess competencies targeted by MR tutoring.

Data Analysis

We performed two types of analyses to determine whether participation in MR raises the mathematics achievement of low-performing first-grade students (Research Question 1) and whether students maintain the gains made in first grade through the end of second grade (Research Question 2). First, we compared the performance of students who had received tutoring (the treatment group) to students who remained on the wait list at the end of first grade (the control group) across the proximal outcome measures (MR1.1 and the MRp) and the more distal WJ AP, QC, and MF subscales. In order to account for the clustering of students within tutors, we conducted these analyses using a two-level hierarchical linear model (HLM). This modeling strategy allowed us to control for student characteristics often associated with variation in early mathematics achievement, including pretest score, gender (female = 1, male = 0), minority status, whether the student has been classified as having a disability (IEP), limited English proficiency, socioeconomic status (measured as whether the student was receiving a free or reduced price lunch), site of treatment (Site 1 comprises 10 schools in a suburban district in State B; Site 2 comprises five schools in an urban district in State B; and Site 3 comprises five schools in three rural districts in State A), and year that the student entered the study (Year 2 = 1, Year 1 = 0). The pretest measure was calculated as the first principal components of participants' scores on the MRp and each of the three WJ III subscales at the start of first grade. We included these explanatory variables in order to hold constant these characteristics across the treatment and control groups as well as to increase the precision through which we measured the effect of participating

in MR. For students whose free and reduced price lunch status (n = 1) or minority status (n = 3) was unavailable from their district, we coded the variable as zero and included an additional indicator variable at Level 1 to flag the observations. The models were:

$$\begin{aligned}\text{Outcome}_{ij} = {} & \beta_{0j} + \beta_{1j}\,(\text{treatment})_{ij} + \beta_{2j}\,(\text{pretest})_{ij} \\ & + \beta_{3j}\,(\text{female})_{ij} + \beta_{4j}\,(\text{Black})_{ij} + \beta_{5j}\,(\text{Hispanic})_{ij} \\ & + \beta_{6j}\,(\text{Other ethnicity})_{ij} + \beta_{7j}\,(\text{FRPL})_{ij} + \beta_{8j}\,(\text{LEP})_{ij} \\ & + \beta_{9j}\,(\text{IEP})_{ij} + \beta_{10j}\,(\text{age at pretest})_{ij} + \beta_{11j}\,(\text{Year2})_{ij} \\ & + \gamma_{00j}\,(\text{Site 1})_{j} + \gamma_{01j}\,(\text{Site 3})_{j} + u_{j} + r_{ij},\end{aligned} \tag{1}$$

where student i is nested within tutor j. The coefficient on treatment, $\beta_{1j,}$ can be interpreted as the average treatment effect. To calculate this as an effect size, we divide β_{1j} by the pooled standard deviation of the outcome being analyzed (first pooled across the treatment and control groups associated with tutors and then pooled across tutors). We also tested for interactions between treatment and year, the site of treatment, and the pretest score. The treatment by year interaction allows us to determine whether the tutors became more effective in their second year of tutoring, as anticipated by the program developers. The treatment by site interactions allowed us to assess whether the effectiveness of tutoring varied across the two main sites where tutors were trained (which also corresponded to the states where the schools were located) and districts classified as suburban (Site 1), urban (Site 2), and rural (Site 3). The treatment by pretest interaction allowed us to determine whether lower performing beginning first graders appear to benefit more from Mathematics Recovery than their higher performing counterparts. This analysis is particularly important, as our study population had a higher than anticipated average pretest score on the WJIII subtests. As a check on these predictions, we also ran the aforementioned HLM model with a sample restricted to those students whose pretest scores were less than a 90 scale score (i.e., students below the 25th percentile nationally) on the Mathematics Reasoning domain of the WJIII (which combines scores on the WJ AP and WJ QC subtests).

Students who participated in the first year of the study (approximately half of our sample) were assessed at the end of second grade on the three WJIII subtests and on more proximal MRp2. The models run to address our second research question were similar to those for first grade, without the control for year.

Our research design also allowed us to describe and compare the growth trajectories of treatment and control cohorts across the first- and second-grade years, punctuated at the end of each 12-week tutoring cycle by the students completing MR tutoring. For example, we used the estimated growth rate of Cohorts 1B and 1C prior to receiving treatment, as well as the estimated growth rates of students on the wait list who did not receive the MR intervention, to estimate a counterfactual to the growth rate of MR

participants in Cohort 1A. At the end of each of these intervals, a given study participant had one of three statuses: not yet received any MR, just completed MR tutoring, or was post MR. We made comparisons within this scheme to determine mathematics achievement outcomes immediately after completing MR tutoring relative to students who had not received the MR intervention (to test the treatment effect) and 12 or 24 weeks after completing MR relative to students who had just completed MR (to test whether MR gains are maintained after the end of treatment).

To estimate these growth trajectories, we used a three-level hierarchical linear growth model (Raudenbush & Bryk, 2002; Singer & Willett, 2002) with repeated observations of WJ III scores indexed at Level 1 by time (indexed by the number of weeks since taking the pretests), time squared (to allow for increasing or decreasing learning rates over first and second grade), time since starting MR, and time since completing MR; at Level 2 by student level demographics; and at Level 3 by school characteristics (e.g., site). To assess whether gains made in MR tutoring were maintained after the tutoring was completed, we constructed a time-varying covariate (POSTMRTIME_{ijt}) that counted the number of weeks after a student completed MR. The Level 1 equation was:

$$\begin{aligned}\text{WJIII}_{ijt} = {} & \pi_{0ij} + \pi_{1ij}(\text{Time})_{ijt} + \pi_{2ij}(\text{Time}^2)_{ijt} \\ & + \pi_{3ij}(\text{MRTIME})_{ijt} + \pi_{4ij}(\text{POSTMRTIME})_{ijt} + \epsilon_{ijt}.\end{aligned} \tag{2}$$

The coefficient π_{3ij} on MRTIME_{ijt} can be interpreted as the weekly average treatment effect—the additional weekly learning associated with participation in MR relative to that of students on the waitlist and in cohorts that have not yet begun tutoring. The coefficient π_{4ij} on POSTMRTIME_{ijt} can be interpreted as the additional weekly learning for participants after completing MR compared to their rate of learning when participating in MR tutoring.

We assumed that the initial status of mathematics achievement, the growth rate of mathematics achievement, and the effect of MR on student achievement varied across students. We assumed that the POSTMRTIME effect was nonrandomly varying. To improve the precision of our estimates, we included student-level, time-invariant covariates that prior research has shown to be associated with student math achievement at the start of first grade, including gender, minority status, whether the student has been classified as having a disability, limited English proficiency, socioeconomic status (measures as whether the student is receiving a free or reduced price lunch), and their age at pretest.

$$\begin{aligned}\pi_{0ij} = {} & \beta_{00j} + \beta_{01j}(\text{Female})_{ij} + \beta_{02j}(\text{Black})_{ij} + \beta_{03j}(\text{Hispanic})_{ij} \\ & + \beta_{04j}(\text{IEP})_{ij} + \beta_{05j}(\text{LEP})_{ij} + \beta_{06j}(\text{frpl})_{ij} + \beta_{07j}(\text{age at pretest})_{ij} \\ & + r_{0ij}; \\ \pi_{1ij} = {} & \beta_{10j} + r_{1ij};\ \pi_{2ij} = \beta_{20j};\ \pi_{3ij} = \beta_{30j} + r_{3ij};\ \pi_{4ij} = \beta_{40j}.\end{aligned} \tag{3}$$

Finally, we assumed that the tutor/school average initial status of mathematics achievement (as measured by the WJ III), tutor/school average weekly growth in mathematics achievement, and the effect of MR on student achievement varied across tutors/schools. We also controlled for site at Level 3 (equation not shown).

WJIII subtests or MRp are missing for less than 1% of study participants at the end of first grade. We treated these data as missing completely at random. Four cases with missing district data on race/ethnicity or receipt of free or reduced price lunch were coded as zero with a missing flag added to each analysis

Results

Study Population

Of our study population, 99% had MR 1.1 diagnostic scores of 11 or below, which is the cut point that the screeners were expected to use to determine if students were eligible for MR. Pretest results on the WJIII instruments indicated that our study population had greater initial mathematics achievement higher than the target population that the developers of MR claim benefit most from MR participation—more than two thirds of the students' scores were above the 25th percentile. As the MR screening process has never been validated against a nationally normed assessment, we do not know if our study population had higher mathematics proficiency at the start of first grade than a business as usual implementation of MR.[5] Nonetheless, as MR has been implemented as a response to intervention (RTI), we paid special attention both to interactions between pretest scores on the effects of participation in MR and to separately estimated treatment effects for students below the 25th percentile on the WJ MR Subscale.

Fidelity of Implementation

We found fidelity to most structural features of implementation to be high. In line with the MR program model, the MR 1.1 diagnostic assessment was conducted with all students at the outset of their tutoring cycles and nearly all tutoring sessions were video-recorded to aid tutor reflection and planning. In addition, the average length of tutoring sessions observed for the fidelity assessment met program expectations and tutors used official MR materials and posed tasks of the kind described in the program model. Given the limited number of evaluation studies of diagnostic interventions and the lack of process-focused measures of implementation fidelity in the literature (O'Donnell, 2008), it is not possible to benchmark the results of our assessment of fidelity to process aspects of the program (i.e., tutors' fidelity to model expectations for the content and nature of tutoring) against findings from previous research. We therefore compared study tutors' levels

of fidelity to either (a) those of expert tutors by coding MR training videos and using the same fidelity coding scheme or (b) levels observed in previous MR implementations (other than those for our evaluation).

As indicated in Table 4, we found that on average, tutors' administration of the MR 1.1 Assessment adhered to program expectations. However, the Learning Framework profiles that tutors assigned at the conclusion of MR1.1 agreed with our coders' assessments slightly less than 70% of the time. This level of fidelity could indicate that tutors had difficulty in using the Learning Framework as intended.

Regarding instructional sessions, we found that tutors' practices were, on average, generally comparable to those of the expert tutors. Although our study tutors were more likely than experts to elicit particular behaviors or directly demonstrate problem-solving strategies, they refrained from making these prohibited, though common, teaching moves 86% of the time. After posing tasks that were too difficult, study tutors were more likely to reduce the difficulty of the task (a move consistent with the MR model) than were expert tutors, who more often directly released students from solving the task. One aspect on which study tutors were more aligned with program expectations than expert tutors was whether they balanced solicitations of students' problem-solving strategies equally across correct and incorrect answers. As shown in Table 4, study tutors' solicitations were significantly more balanced (around 0). However, we found that study tutors fell short of program expectations on two measures of exposure. First, the median number of lessons delivered, 32, was significantly less than the stated program expectation of 48 to 60 lessons and the "operational expectation" of 42 lessons based on data collected by the U.S. Math Recovery Council. Second, tutors spent less time per session on strategy-based activities (and more time on number words and numerals) than program developers expect. However, only the amount of time less than program expectations spent on strategy-based activities per session was significantly related to students' progress on the MR1.1, suggesting that the type of activities on which lessons focus is more important than the number of lessons delivered (Munter, 2010).

Both of these exposure measures—average number of tutoring sessions per student and the average number of minutes spent on strategy-based activities per session—were, generally, unrelated to outcomes at the end of Year 2. The exception was a statistically significant relationship between WJ MF scores and a variable that weighted the number of lessons by average strategy-based time. Only one process-oriented measure of fidelity—the ratio of the frequency with which tutors elicited particular behaviors to the frequency with which they solicited students' strategies—emerged as consistent predictor of end of Year 2 outcomes. Surprisingly, however, this behavior eliciting ratio was not related to most end of Year 1 outcomes. Behavior eliciting is prohibited (because it may lead to unproductive "guessing

Table 4
Comparison of Fidelity Indexes

Fidelity Index	Expert Mean	Study Mean
Average rate with which tutors administered the MR 1.1 diagnostic assessment correctly, not changing the nature of the assessment	n/a	0.93
Average percentage of Math Recovery (MR) Learning Framework aspects for which tutors used the 1.1 assessment to generate sufficient information to assign a profile stage/level	n/a	0.84
Average percentage of MR Learning Framework aspects on which the tutor's assignment of student profile matches that of the coder	n/a	0.69
Percentage of tasks that were genuine problems (neither too easy nor too difficult)	0.65	0.62
Percentage of tasks for which tutors reduced the level of difficulty when the initial level of task challenge was too difficult	0.39	0.83***
Percentage of tasks for which tutors provided sufficient wait-time	0.93	0.87
Percentage of tasks in which tutors refrained from eliciting particular behaviors	0.93	0.86*
Percentage of tasks for which tutors refrained from directly demonstrating strategy	0.98	0.86***
When, on average, tutors asked student to check answer: 0 represents perfect balance between incorrect (–1) and correct (1) answers	–0.07	0.13***
When, on average, tutors solicited student's strategy (scored as previous index)	0.28	0.07***
	MR model	
Average number of tutoring sessions per student	48–60(42 typical)	32.49***
Average length of tutoring sessions in minutes	25–30	25.04
Average number of minutes spent on strategy-based activities per session	10–13.5 (40%-45% of session)	6.54*** (26% of session)

*$p < .05$. ***$p < .001$ (two-tailed tests).

games"), and soliciting students' strategies (after both correct and incorrect responses) is encouraged by the MR model. Our analysis of the relative frequencies of these moves was not motivated directly by MR expectations but instead was based on a conjecture that more direct intervention on the

tutor's part than specified by the MR model may have less of a negative impact if balanced by moves that place the responsibility for explaining solutions on students. Thus, our finding points to a slightly different message than that promoted by MR.

Because this was an effectiveness study (and not an efficacy trial), we did not intervene to influence tutors' practices (O'Donnell, 2008), but instead assessed the implementation of the program in schools implementing MR for the first time, including how tutors implemented the program based on the training they had received from MR personnel. The findings of the fidelity assessment point to areas of the MR model that are likely challenging to implement as intended, but also indicate that the implementation of MR in our evaluation was aligned with program guidelines sufficiently to attribute study outcomes to the MR program with confidence.

Outcomes

We examined the effects of receiving MR tutoring at the end of first grade by combining Year 1 and Year 2 participants. The results reveal large effects on the MR 1.1 Assessment (effect size 1.04), the assessment in which the tasks are similar to those used in tutoring sessions, and small to moderate effects (effect sizes of .15–.30) on the MRp proximal measure and the more distal measures of the three WJ III subtests (HLM regression tables are included in the Appendix Tables A1-A6 available in the online version of the journal). Specifically, differences in the effects of treatment on the WJ III subtests at the end of first grade between students who had received tutoring and those on the waitlist were .24 on the quantitative concepts scale and .28 on the applied problems scale, with an effect size .30 on the combined Math Reasoning subscale (all differences statistically significant at the $p < .001$ level). In other words, participation in MR was associated with a 30% of a standard deviation increase in Math Reasoning Score. Unexpectedly, the effects of receiving MR tutoring on the MRp (.29) were closer to the WJIII subtest scores than the MR 1.1 Assessment. The effect size on the WJIII Math Fluency assessment was small (.15). Effect sizes were larger when the sample was restricted to students scoring at or below the 25th percentile on the WJ MR Subscale: 1.14 on the MR Initial Assessment and between 0.3 and 0.4 on MRp and the WJIII subscales. Effect sizes are listed in Table 5.

We found no significant interaction between treatment and year, indicating that the MR tutors in our study were not measurably more effective in their second year of tutoring than in their first—a result that is corroborated by findings from our fidelity assessment and that contradicts the expectations of the MR developers. This suggests that the costs of tutor turnover come mainly from the expense of training new tutors, and that little may be lost if the role of tutor is rotated periodically among teachers in the

Table 5
Effect Sizes at the End of First Grade for Math Recovery (MR) Participation Calculated From the Hierarchical Linear Modeling (HLM) Models in Tables A1-A6 (in the online version of the journal)

	Full Sample		Sample With WJIII MR Score Less Than 90	
Assessment	Effect Size	*N*	Effect size	*N*
MR1.1 (Diagnostic Assessment)	1.04***	759	1.15***	229
MR Proximal (MRp)	0.29***	775	0.32**	235
WJIII Math Fluency (MF)	0.15*	775	0.31**	235
WJIII Applied Problems (AP)	0.28***	775	0.38**	235
WJIII Quantitative Concepts (QC)	0.24***	775	0.30**	235
WJIII Math Reasoning (AP + QC)	0.30***	775	0.40*	235

Note. A scale score of 90 on the Woodcock Johnson III (WJIII) Math Reasoning Scale marks the 25th percentile for a nationally normed sample.
*$p < .05$. **$p < .01$. ***$p < .001$ (two tailed tests).

school. We found a significant treatment by site interaction on the MR1.1, suggesting that MR was implemented with less effectiveness at the site with five rural schools. However, this site effect was not evident in other assessments

We also found a significant interaction between students' performance on the pretest and participation in MR on some of the assessments, indicating that students with lower skills at the start of first grade are likely to benefit more from MR tutoring. Figure 2a shows the predicted values on the Math Reasoning subscale of the WJIII for study participants at the end of first grade by percentile of performance on the pretest measure and whether they received MR tutoring (treatment) or remained on the waitlist (control) at the end of the year. Figure 2b shows the corresponding effect sizes. Students receiving MR tutoring have predicted *w* scores of about 4 points higher than their counterparts who did not receive MR tutoring. This translates to a predicted effect size between .4 and .5 for those in the lowest quartile of study participants at pretest (see Figure 2b). As pretest scores increase, the predicted effect sizes converge on zero. As the number of days of tutoring ranged from 3 to 52, we tested the sensitivity of our results to the amount of tutoring received. To do this, the "treatment" variable in the models described previously was replaced with the number of MR lessons received. Among those receiving tutoring, 10% received 23 or fewer sessions and 10% (the 90th percentile) received 43 or more sessions (the median was 32). Students who remained on the waitlist at the end of the year were coded as zero. The square of the number of sessions received was also included

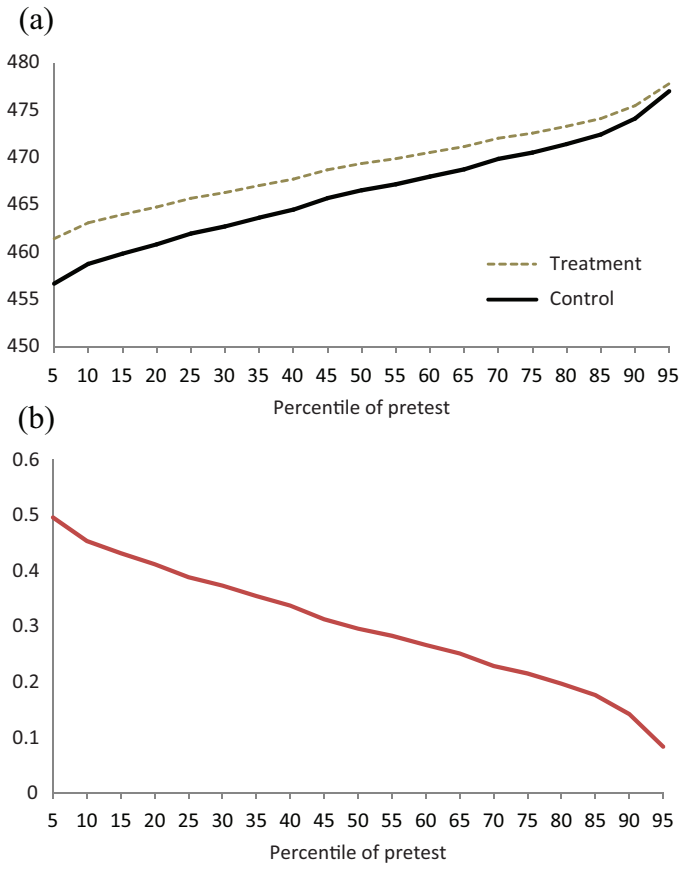


Figure 2. **a. Predicted values on the Woodcock Johnson III (WJIII) Mathematical Reasoning Subscale at the end of first grade, by percentile of score on pretest and participation in Math Recovery (MR) or the control group. b. Predicted effect sizes at the end of first grade of participation in MR on the WJIII Mathematical Reasoning Subscale, by percentile of score on pretest**

in the models to test whether the effect of participating in MR decreases with the number of lessons received. The results of the HLM models run for each assessment are presented in Table A7 in the online version of the journal. Number of lessons had a small, positive, and statistically significant coefficient in each of the models except the one predicting variation in WJ MF; the quadratic term was negative and statistically significant for each model

Table 6

Predicted Assessment Scores at the End First Grade for Students Receiving 23 (10th percentile), 32 (median), and 43 (90th percentile) Math Recovery (MR) Lessons (based on hierarchical linear modeling models in Table A7, in the online version of the journal)

Assessment	23 Lessons	32 Lessons	43 Lessons
MR1.1 (Diagnostic Assessment)	17.6	17.8	17.2
MR Proximal (MRp)	7.4	7.5	7.4
WJIII Math Fluency (MF)	485.5	485.5	485.4
WJIII Applied Problems (AP)	463.9	464.2	463.7
WJIII Quantitative Concepts (QC)	474.8	474.4	472.7
WJIII Math Reasoning (AP + QC)	469.3	469.2	468.1

Table 7

Effect Sizes at the End of Second Grade for Math Recovery (MR) Participation Calculated From the Hierarchical Linear Modeling Models in Table A7 (in the online version of the journal)

Assessment	Effect size	Significance	*N*
MR Second Grade Proximal Assessment (MRp2)	0.07	*ns*	320
MR Proximal (MRp)	–0.02	*ns*	320
WJIII Math Fluency (MF)	0.09	*ns*	320
WJIII Applied Problems (AP)	0.00	*ns*	320
WJIII Quantitative Concepts (QC)	0.06	*ns*	320

*$p < .05$. **$p < .01$. ***$p < .001$ (two tailed tests).

except those predicting WJ MF and WJ AP. To illustrate the magnitude of this relationship, we calculated the predicted scores for each assessment for students at the 10th percentile (23), median (32), and 90th percentile (43) of MR lessons received (results shown in Table 6). The predicted w score on the WJ MR subscale of the WJII was 469.3 for students receiving 23 MR lessons, 469.2 for students receiving 32 MR lessons, and 468.1 for students receiving 43 MR lessons. In contrast, the number of lessons is positively and significantly related to only the WJ MF outcome when the sample is restricted to the 107 students included in our fidelity assessment. Across both samples, though, the effect of additional lessons was small.

To examine whether the positive effects of participating in MR were maintained, we assessed the Year 1 participants at the end of second grade on the WJ MF, AP, and QC subtests, as well as on the second-grade version of the MR Proximal Assessment (MRp2). As shown in Table 7, across each of these assessments the treatment effects at the end of second grade were small and none were statistically significant. Interactions between pretest

and treatment were also small and not statistically significant (results not shown). These results suggest that the positive impact of MR seen at the end of first grade fades by the end of second grade.

Results from our growth model analysis highlight the relative size of the gains from participation in MR in first grade relative to students on the waitlist or who have not yet begun their MR tutoring (the control group at a particular point in time), as well as the dissipation of these gains by the end of second grade. Appendix Table 9 shows increases in mathematics achievement for MR participants across all WJIII subscales during the tutoring period (with $p < .05$ in each case). However, this growth rate tends not to maintain after completion of MR. For example, on the WJ AP, MR participants gained 1.08 points per week, on average, during tutoring while students who had not begun tutoring or were on the waitlist for MR gained 0.83 points per week across the same time period. After exiting MR tutoring, however, participants' growth trajectories reverted back to pretutoring rates—a rate of 0.79 per week on the WJ AP. Figures 3a, 3b, and 3c show the predicted growth rates for members of MR Cohort A, B, and C, respectively, relative to counterfactual, their expected growth rate in the absence of treatment. The counterfactual is estimated based on the performance of students on the waitlist for MR as well as the performance of students whose cohort has not yet begun tutoring. The pattern of gains while receiving tutoring, with regression toward the growth trajectories of nonparticipants, was consistent across assessments. The predicted values of participants' scores at the end of second grade are similar for treatment and control participants.

Discussion

The findings of this study have implications for practice and research. Practically, the positive causal effect of MR tutoring at the end of first grade demonstrates that tutoring programs that are diagnostic rather than scripted in nature can overcome fidelity concerns and have a positive impact on students' mathematics performance. The end of first grade effect sizes are broadly comparable to those in interventions reviewed by Slavin and Lake (2006), including several cooperative learning programs that had median effect sizes of at least +0.30 in studies using randomized experimental or randomized quasi-experimental designs. The predicted effect sizes for students scoring below the 25th percentile on the math reasoning subscale of the WJIII at the beginning of first grade are also practically significant, with student gains of between 30% and 40% of a *SD* on externally administered, nationally normed assessments.

The MR effects were also broadly similar to those found in Ritter, Barnett, Denny, and Albin's (2009) meta-analysis of randomized field trials of elementary and middle school volunteer tutoring programs, where the resulting effect sizes were mixed and averaged to a non–statistically significant

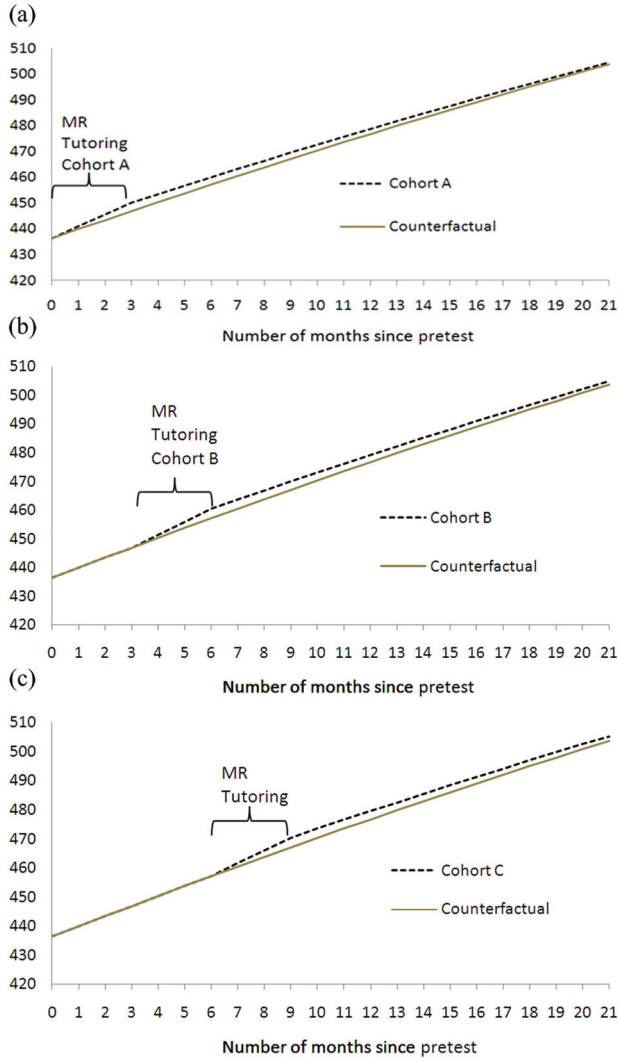


Figure 3. **a. Predicted values on the Woodcock Johnson III (WJIII) Applied Problems Subtest, by months since taking the pretest and participation Math Recovery (MR) in Cohort A and the counterfactual. b. Predicted values on the WJIII Applied Problems Subtest, by months since taking the pretest and participation MR in Cohort B and the counterfactual. c. Predicted values on the WJIII Applied Problems Subtest, by months since taking the pretest and participation MR in Cohort C and the counterfactual.**

.27. A meta-analysis of 52 studies on the relationship between tutoring and student achievement (Cohen et al., 1982) found average effect sizes greater than .40—higher than the effects of MR on the MRp and the WJ III measures but lower than effects on the MR1.1.

The magnitude of the difference in impacts between the tutor-administered MR1.1 and the other assessments we used (i.e., the MRp and the WJIII subtests) raises practical concerns for the MR program. The MR1.1, on which the largest effects were achieved, was administered by the tutor rather than an external assessor and involved types of tasks and materials with which children had become familiar in tutoring sessions. The results are of similar magnitude to two previous quasi-experimental studies of MR that showed large effect sizes on an in-house measure of students' arithmetical competence, the MR Assessment in Early Numeracy, which was specific to the context of MR tutoring (MacLean, 2003; Phillips et al., 2003). Large effects of tutoring on the MR1.1 suggest why this program has been adopted relatively widely: Tutors are likely to observe that the children with whom they are working are making significant progress during tutoring sessions. On nationally normed assessments, however, the gains are more modest and appear to fade by the end of second grade.

With regard to practice, our findings have little to say about the efficacy of either one-on-one tutoring interventions or inquiry-oriented instructional approaches in general. The differential impacts on the MR1.1 when compared with the other assessments we used suggest that MR tutoring may need to bridge more systematically to regular classroom instruction if the program is to achieve its primary goal of enabling students to be successful in the context of regular classroom instruction. Other intervention research suggests that classroom instruction that does not take account of what children already know could result in students not benefiting long term from intensive, relatively short-term interventions (Clements, Sarama, Spitler, Wolfe, & Lange, 2011; Sarama, Clements, Wolfe, & Spitler, 2011). The tutors in our study were not allowed to collaborate with first- and second-grade teachers in their schools in order to ensure that MR instructional approaches were not used with control students in their classrooms. The MR developers report that in regular implementations of MR, the intervention teachers can become important sources of expertise for classroom teachers and provide ongoing support in the classroom for discontinued MR students. Our findings suggest that support of this type may need to be more than incidental. Additional work is needed to explore the plausibility of this explanation. However, given the small to medium effect sizes that we found on the WJIII measures at the end of first grade and the fade-out of effects by the end of second grade, we suggest that school districts carefully balance the costs of MR with these potential benefits.

Further research is needed to understand why initial gains made by participants diminish after tutoring ends. This concern also plagues Reading Recovery—a program with a broadly similar protocol. The developer of

Reading Recovery proposed that schools should identify low-achieving children early in their school careers and provide a clinical intervention program that supports their development of strategies on which they can subsequently build in the context of classroom instruction (Hurry & Sylva, 2007). This proposition has not been upheld for Reading Recovery (D'Agostino & Murphy, 2004), and our study suggests that it might not hold up for Math Recovery either. The fade out of the effects of these interventions calls for more research to clarify the reasons why children are not developing capabilities that can support their subsequent classroom learning. Possible explanations range from neurocognitive differences to differences in out-of-school learning opportunities. In addition, it will be important to investigate whether lack of follow-up that is coordinated with classroom instruction contributes to the lack of sustained gains. If this latter possibility proves to be the case, it would call into question the core assumptions on which both Mathematics Recovery and Reading Recovery are based and indicate the need for ongoing support throughout the elementary grades for at least some students who are behind at the beginning of first grade.

Notes

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305B070554 and R305B040110 to Vanderbilt University. The opinions expressed are those of the authors and do not necessarily represent views of the U.S. Department of Education. Paul Cobb had a potential conflict of interest at the time this study was conducted in that his spouse had a long-term contract with the U.S. Math Recovery Council. An external reviewer commissioned by Vanderbilt reviewed multiple iterations of this manuscript for bias.

[1]Of the 360 students selected at the beginning of Years 1 and 2 for participation in Math Recovery (MR), 83% completed treatment, 9% withdrew from school, 3% were placed in special education, and each of 1% or less were either determined no longer eligible for treatment by MR staff, exited by MR staff because of excessive absences, began receiving English language learner (ELL) math support from an MR trained teacher, or, in 2 cases, treatment was stopped without study approval because the MR tutor felt they were making a "lack of sufficient progress." Of the 562 assigned to the waitlist, 80% were still on the waitlist at the end of the study, 8% had moved into a treatment cohort, 10% had withdrawn from the treatment school, and the remaining 2% had extended absences or were assigned to receive special education or other math supports.

[2]While we carefully studied the implementation of MR tutoring (e.g., the organization and content of the tutoring sessions), we did not assess the quality or fidelity to the MR model of the initial or ongoing training that the MR tutors in our study received. This was mainly because the training was delivered by the same members of the U.S. Math Recovery Council that we worked with to design the framework for assessing fidelity and who trained our fidelity coders on MR.

[3]We have no internal consistency information on the MR 1.1 Diagnostic Assessment, although it is correlated with the Woodcock Johnson III (WJIII) Math Fluency (WJ MF) assessment at .28 and the WJIII Mathematics Core Reasoning (WJ MR) subscore at .50. Test-retest reliability between fall and winter assessments among students on waitlist or in cohorts that had not yet started tutoring was .622.

[4]At the start of first grade, the MR proximal assessment (MRp) is correlated at .33 with the WJIII MF and .44 with the WJIII MR subscore. An item response theory (IRT) analysis conducted on the MRp indicated good internal consistency (Kuder-Richardson 20/21 coefficient = .759). There is evidence of high construct validity from documentation of the

"Four Building Blocks" (Wilson, 2005) and the internal structure of the instrument (Spearman's Rho = .93). All students received the same form at pretest (form A) and were randomly assigned forms B or C at the end of Cohort 1, students who took form B at the end of Cohort 1 took form C at the end of Cohort 2, and vice versa. All students then took form A again at the end of Cohort 3 (the end of first grade). There were no statistically significant differences between the scores of Year 1 students who took forms B or C at either the end of Cohort 1 or the end of Cohort 3.

[5]In addition, data collected on students tutored between 2004 and 2006 by the U.S. Math Recovery Council indicate that 85% of tutored students had met the screening criteria implemented in our study and that 15% had higher scores on the screening instrument. In addition, we were told by MR representatives that tutoring cohorts in business as usual implementations of MR are not populated starting with students that have the lowest MR diagnostic scores, rather they are chosen from throughout the distribution of students identified as eligible so that the tutor is working with students at different levels. Thus, drawing randomly from the eligible population in our study does not appear to differ so much from business as usual (BAU), at least for Cohort 1.

References

Alexander, K. L., & Entwistle, D. R. (1998). Achievement in the first two years of school: Patterns and processes. *Monographs of the Society for Research in Child Development*, *53*(Serial No. 218).

Aubrey, C., Dahl, S., & Godfrey, R. (2006). Early mathematics development and later achievement: Further evidence. *Mathematics Education Research Journal*, *18*(1), 27–46.

Aunola, K., Leskinen, E., Lerkkanen, M. K., & Nurmi, J. E. (2004). Developmental dynamics of math performance from preschool to grade 2. *Journal of Educational Psychology*, *94*(4), 699–713.

Baroody, A. J. (1987). The development of counting strategies for single-digit addition. *Journal for Research in Mathematics Education*, *18*, 141–157.

Baroody, A. J. (1990). How and when should place-value concepts and skills be taught? *Journal for Research in Mathematics Education*, *21*, 281–286.

Baroody, A. J., & Ginsburg, H. P. (1986). The relationship between initial meaningful and mechanical knowledge of arithmetic. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 75–112). Hillsdale, NJ: Lawrence Erlbaum.

Baroody, A. J., & Tiilikainen, S. H. (2003). Two perspectives on addition development. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills: Constructing adaptive expertise* (pp. 75–125). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Barton, P. E. (2003). *Parsing the achievement gap: Baselines for tracking progress*. Princeton, NJ: ETS

Beishuizen, M. (1993). Mental strategies and materials or models for addition and subtraction up to 100 in Dutch second grades. *Journal for Research in Mathematics Education*, *24*, 294–323.

Berends, M., Lucas, S. R., Sullivan, T., & Briggs, R. J. (2005). *Examining gaps in mathematics achievement among racial-ethnic groups, 1972–1992*. Santa Monica, CA: RAND Corp.

Burch, P. (2007). *Supplemental education services under NCLB: Emerging evidence and policy issues*. East Lansing, MI: The Great Lakes Center for Education Research & Practice.

Carpenter, T. P., Franke, M. L., Jacobs, V. R., Fennema, E., & Empson, S. B. (1997). A longitudinal study of invention and understanding in children's multidigit addition and subtraction. *Journal for Research in Mathematics Education*, *29*, 3–20.

Carpenter, T. P., & Moser, J .M. (1982). The development of addition and subtraction problem solving skills. In T. P. Carpenter, J. M. Moser, & T. A. Romberg (Eds.), *Addition and subtraction: A cognitive perspective* (pp. 9–24). Hillsdale, NJ: Lawrence Erlbaum.

Carpenter, T. P., & Moser, J. M. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal for Research in Mathematics Education*, *15*, 179–202.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, *25*, 471–533.

Clements, D. H. (1999). Subitizing: What is it? Why teach it? *Teaching Children Mathematics*, *5*, 400–405.

Clements, D. H., & Sarama, J. (2007). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education*, *38*(2), 136–163.

Clements, D. H., & Sarama, J. (2012). Early childhood mathematics intervention. *Science*, *333*, 968–970.

Clements, D. H., Sarama, J., Spitler, M. E., Wolfe, C. B., & Lange, A. A. (2011, March). Longitudinal evaluation of a scale-up model for teaching math with trajectories and technologies: Persistence of effect in the third year. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness, Washington, DC.

Cobb, P., Gravemeijer, K., Yackel, E., McClain, K., & Whitenack, J. (1997). Mathematizing and symbolizing: The emergence of chains of signification in one first-grade classroom. In D. Kirshner & J. A. Whitson (Eds.), *Situated cognition theory: Social, semiotic, and neurological perspectives* (pp. 151–233). Mahwah, NJ: Lawrence Erlbaum.

Cockcroft, W. (1982). *Mathematics counts*. London: HMSO.

Cohen, P. A., Kulik, J. A., & Kulik, C. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, *19*, 237–248.

Common Core State Standards Initiative. (2010). *Common core state standards for mathematics*. Washington, DC: National Governors Association Center for Best Practices, Council of Chief State School Officers.

Crosnoe, R. (2005). Double disadvantage or signs of resilience? The elementary school contexts of children from Mexican immigrant families. *American Educational Research Journal*, *42*(2), 269–303.

D'Agostino, J., & Murphy, J. (2004). A meta-analysis of Reading Recovery in United States schools. *Educational Evaluation and Policy Analysis*, *26*, 23–38.

Derry, S. J., & Potts, M. K. (1998). How tutors model students: A study of personal constructs in adaptive tutoring. *American Educational Research Journal*, *35*, 65–99.

Duncan, G. J., Claessens, A., & Engel, M. (2004). *The contribution of hard skills and socio-emotional behavior to school readiness*. Retrieved from http://www.northwestern.edu/ipr/people/duncanpapers.html.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research*, *28*, 237–256.

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.

Fuson, K. C. (1988). *Children's counting and concepts of number*. New York, NY: Springer-Verlag.

Fuson, K. C. (1992). Learning addition and subtraction: Effects of number words and other cultural tools. In J. Bideaud, C. Meljac, & J. P. Fischer (Eds.), *Pathways to number: Children's developing numerical abilities* (pp. 283–306). Hillsdale, NJ: Lawrence Erlbaum.

Fuson, K. C., Wearne, D., Hiebert, J., Human, P., Olivier, A., Carpenter, T., . . . Fennema, E. (1997). Children's conceptual structure for multidigit numbers and methods of multidigit addition and subtraction. *Journal for Research in Mathematics Education*, *28*, 130–162.

Griffin, S., & Case, R. (1999). Re-thinking the primary school math curriculum: An approach based on cognitive science. *Issues in Education*, *3*(1), 1–49.

Hiebert, J., & Wearne, D. (1992). Links between teaching and learning place value with understanding in first grade. *Journal for Research in Mathematics Education*, *23*, 98–122.

Houssart, J. (2001). Counting difficulties at Key Stage 2. *Support for Learning, 16*, 11–16.

Hurry, J., & Sylva, K. (2007). Long term outcomes of early reading intervention. *Journal of Research in Reading, 30*, 227–248.

Keogh, B. K., Major, S. M., Reid, H. P., Gandara, P., & Omori, H. (1978). Marker variables: A search of comparability and generalizability in the field of learning disabilities. *Learning Disability Quarterly*, *1*(Summer), 5–11.

Lesh, R., & Kelly, A. E. (1997). Teachers' evolving conceptions of one-to-one tutoring: A three-tiered teaching experiment. *Journal for Research in Mathematics Education, 28*, 398–430.

MacLean, H. E. (2003). *The effects of early intervention on the mathematical achievement of low-performing first grade students* (Unpublished doctoral dissertation). University of Houston, Houston, TX.

McArthur, D., Stasz, C., & Zmuidzinas, M. (1990). Tutoring techniques in algebra. *Cognition and Instruction*, *7*, 197–244.

McWayne, C. M., Fantuzzo, J. W., & McDermott, P. A. (2004). Preschool competency in context: An investigation of the unique contribution of child competencies to early academic success. *Developmental Psychology*, *40*(4), 633–645.

Mosher, F. A. (2011). *The role of learning progressions in standards-based education reform* (CPRE Policy Brief). Philadelphia, PA: CPRE.

Munter, C. (2010). *Evaluating Math Recovery: The impact of implementation fidelity on student outcomes* (Unpublished doctoral dissertation). Vanderbilt University, Nashville, TN.

Oakes, J. (1990). *Multiplying inequalities: The effect of race, social class, and tracking on opportunities to learn mathematics and science*. Santa Monica, CA: The RAND Corporation.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research*, *78*(1), 33–84.

Phillips, V. J., Leonard, W. H., Horton, R. M., Wright, R. J., & Stafford, A. K. (2003). Can Math Recovery save children before they fail? *Teaching Children Mathematics*, *10*(2), 107–111.

Princiotta, D., Flanagan, K. D., & Germino Hausken, E. (2006). *Fifth grade: Findings from the fifth-grade follow-up of the early childhood longitudinal study, kindergarten class of 1998-99 (ECLS-K)* (NCES 2006-038). Washington, DC: National Center for Education Statistics.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.

Ritter, G. W., Barnett, J. H., Denny, G. S., & Albin, G. R. (2009). The effectiveness of volunteer tutoring programs for elementary and middle school students: A meta-analysis. *Review of Educational Research*, *79*(3), 3–38.

Rolfhus, E., Gersten, R., Clarke, B., Decker, L., Wilkins, C., & Dimino, J. (2012). *An evaluation of Number Rockets: A Tier 2 intervention for grade 1 students at risk for difficulties in mathematics* (NCEE 2012-4007). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Sarama, J., Clements, D. H., Wolfe, C. B., & Spitler, M. E. (2011, April). *Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies*. Paper presented at the annual meeting of AERA, New Orleans, LA.

Schmidt, W. H., McKnight, C. C., Cogan, L. S., Jakwerth, P. M., & Houang, R. T. (2002). *Facing the consequences: Using TIMSS for a closer look at the United States mathematics and science education* (eBook ed.). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Singer, J. D., & Willett, J. B. (2002). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York, NY: Oxford University Press.

Slavin, R. E., & Lake, C. (2006). *Effective programs in elementary mathematics: A best-evidence synthesis*. Baltimore, MD: Johns Hopkins University, Center for Data-Driven Reform in Education.

Starkey, P., & Klein, A. (2005). *A longitudinal study of the effects of a pre-kindergarten mathematics curriculum on low-income children's mathematical knowledge* (From PCER 2002: Grantee Annual Progress Report (2005), IES Grant No. R305J020026). Berkeley: University of California.

Steffe, L. P., Cobb, P., & von Glasersfeld, E. (1988). *Construction of arithmetical meanings and strategies*. New York, NY: Springer-Verlag.

Steffe, L. P., von Glasersfeld, E., Richards, J. J., & Cobb, P. (1983). *Children's counting types: Philosophy, theory and application*. New York, NY: Praeger Publishers.

Stevenson, H. W., & Newman, R. S. (1986). Long-term prediction of achievement and attitudes in mathematics and reading. *Child Development*, *57*, 646–659.

Tang, E. P., & Ginsburg, H. P. (1999). Young children's mathematical reasoning: A psychological view. In *Developing mathematical reasining in grades K–12: 1999 yearbook* (pp. 45–61). Reston, VA: NCTM.

Wagner, S. (2005). *PRIME: PRompt Intervention in Mathematics Education: Executive summary of research and programs*. Columbus, OH: Ohio Resource Center for Mathematics, Science, and Reading & Ohio Department of Education.

Wasik, B. A., & Slavin, R. E. (1993). Preventing early reading failure with one-to-one tutoring: A review of five programs. *Reading Research Quarterly*, *28*, 178–200.

Wilms, J. D. (1986). Social class segregation and its relationship to pupils' examination results in Scotland. *American Sociological Review*, *51*, 224–241.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence-Erlbaum.

Wright, R. J. (1991). What number knowledge is possessed by children beginning the kindergarten year of school? *Mathematics Education Research Journal*, *3*(1), 1–16.

Wright, R. J. (1994). A study of the numerical development of 5-year-olds and 6-year-olds. *Educational Studies in Mathematics*, *26*(1), 25–44.

Wright, R. J. (2003). Mathematics Recovery: A program of intervention in early number. *Australian Journal of Learning Disabilities*, *8*(4), 6–11.

Wright, R. J., Martland, J., & Stafford, A. K. (2000). *Early numeracy: Assessment for teaching and intervention*. Thousand Oaks, CA: Sage Publications, Inc.

Wright, R. J., Martland, J., Stafford, A. K., & Stanger, G. (2002). *Teaching Number: Advancing children's skills and strategies*. London: Paul Chapman Publishing.
VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction*, *21*, 209–249.

Manuscript received February 9, 2011
Final revision received October 8, 2012
Accepted November 1, 2012