

An Evaluation of Number Rockets: A Tier-2 Intervention for Grade 1 Students at Risk for Difficulties in Mathematics



An evaluation of *Number Rockets*: a Tier-2 intervention for grade 1 students at risk for difficulties in mathematics

Final Report

February 2012

Authors:

Eric Rolfhus, Ph.D., Edvance Research, Inc.

Russell Gersten, Ph.D., Instructional Research Group

Ben Clarke, Ph.D., University of Oregon

Lauren E. Decker, Ph.D., Edvance Research, Inc.

Chuck Wilkins, Ph.D., Edvance Research, Inc.

Joseph Dimino, Ph.D., Instructional Research Group

Project Officer:

**Karen Armstrong
Institute of Education Sciences**

NCEE 2012-4007

U.S. Department of Education



U.S. Department of Education

Arne Duncan

Secretary

Institute of Education Sciences

John Q. Easton

Director

National Center for Education Evaluation and Regional Assistance

Rebecca A. Maynard

Commissioner

February 2012

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, under contract ED-06C0-0017 with Regional Educational Laboratory Southwest administered by the Edvance Research.

IES evaluation reports present objective information on the conditions of implementation and impacts of the programs being evaluated. IES evaluation reports do not include conclusions or recommendations or views with regard to actions policymakers or practitioners should take in light of the findings in the report.

This report is in the public domain. Authorization to reproduce it in whole or in part is granted. While permission to reprint this publication is not necessary, the citation should read: Rolfhus, E., Gersten, R., Clarke, B., Decker, L., Wilkins, C., and 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.

This report is available on the Institute of Education Sciences website at <http://ncee.ed.gov> and the Regional Educational Laboratory Program website at <http://edlabs.ed.gov>.

Alternate Formats Upon request, this report is available in alternate formats, such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

Disclosure of potential conflict of interest

Dr. Ben Clarke, a co-author of this report, is also a co-author of the Quantity Discrimination subtest (Clarke et al. 2006) that was used as part of the screener in this study. Dr. Clarke did not participate in the screening, data collection, or analysis .

Regional Educational Laboratory Southwest (REL Southwest) contracted with individuals and entities to obtain expert advice and technical assistance for this study. Although these contractors' professional work may not be entirely independent of or separable from the tasks they carried out for REL Southwest, no one from the Instructional Research Group or the University of Oregon who was associated with this study has financial interests that could be affected by the content of this report.¹

¹ Contractors carrying out research and evaluation projects for IES frequently need to obtain expert advice and technical assistance from individuals and entities whose other professional work may not be entirely independent of or separable from the tasks they are carrying out for the IES contractor. Contractors endeavor not to put such individuals or entities in positions in which they could bias the analysis and reporting of results, and their potential conflicts of interest are disclosed.

Acknowledgments

This study was a collaborative effort by many individuals and organizations worthy of acknowledgments. The study authors would like to thank all the district personnel, principals, teachers, and students for their patience during the evaluation and for allowing access to their schools throughout the year. In addition, the participation of the tutors was critical to the success of this study. Without the participation of these individuals, this evaluation would not have been possible.

Through the involvement and guidance of a number of technical experts, the study team was able to clarify and solve many complex technical and analytical questions. The team of experts included David Francis, David Myers, David Chard, and Jeremy Kilpatrick, members of the study's technical working group. Mengli Song from American Institutes for Research also provided timely and expert guidance on a variety of statistical issues. Scott Baker, Alice Klein, and Lynn Fuchs also served in an advisory capacity. In addition, Lynn, along with Joan Bryant and Caitlin Craddock, members of the original *Number Rockets* development team, provided materials and consultation on training and implementation of the intervention.

The study authors would also like to thank the many individuals who contributed significantly to this study. Jessica Brite, research associate, supported the project, was integral to all aspects, and was critical to its success. Dan Hunt, project manager, planned and provided site and logistical management across four states through natural disasters, throughout all phases of the evaluation from the initial study planning to final report preparation. Denise Clyburn, project coordinator, assisted in planning and site-management activities and provided a wide range of assistance throughout the evaluation. David Mellinger provided assistance in scoring and comprehensive database management. Laural Logan-Fain led recruitment efforts during some of the most difficult and challenging situations, such as natural disasters, with remarkable persistence and patience. Meg Grant contributed to the recruitment activities and directed all study dissemination efforts with insightful understanding of constituents. Debrale Graywolf provided comprehensive and essential editorial services and guidance. Lois Gregory, Noelle Howland, and Debora Carmichael of Edvance, along with Becky Newman-Gonchar and Herb Turner of Analytica, Inc., provided expert quality review and invaluable advice throughout report preparation and multiple review cycles. Suellen Bowman, JoAnne Sapp, and Mary Jo Taylor provided specialized training and coaching. Sue Buckley, Terry Buckley, Kelly Haymond, Elaine Livingston, Beverly Vance, Debbie Vignovich, and Carol Wilner conducted pretest and posttest data collection. Additional thanks to Ermine Orta for expert assistance in the planning stages of this study. Also, Derrick Flores and Patrick Guerra supported the technological needs of this study throughout the project.

A final thanks to Don Barfield, project sponsor, who provided strategic leadership throughout the project and guidance in successfully recruiting and working with the study schools.

Contents

DISCLOSURE OF POTENTIAL CONFLICT OF INTEREST	III
ACKNOWLEDGMENTS.....	IV
EXECUTIVE SUMMARY.....	IX
DESCRIPTION OF THE PROGRAM.....	XI
STUDY DESIGN, METHODOLOGY, AND IMPLEMENTATION	XI
ANALYSIS AND FINDINGS	XIII
CONCLUSIONS.....	XIII
LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH.....	XIV
CHAPTER 1: INTRODUCTION AND STUDY OVERVIEW	1
IMPORTANCE OF EARLY INTERVENTION IN MATHEMATICS	2
RESPONSE TO INTERVENTION	3
RESPONSE TO INTERVENTION TIER 2 MATHEMATICS INTERVENTION RESEARCH	4
DESCRIPTION OF THE PROGRAM.....	6
CURRENT STUDY	6
STRUCTURE OF THE REPORT	9
CHAPTER 2: STUDY DESIGN AND METHODOLOGY	10
STUDY DESIGN OVERVIEW	10
STUDY TIMELINE	12
POWER ANALYSIS AND SAMPLE SIZE	15
TARGET DISTRICT POPULATION AND RECRUITMENT PROCESS.....	15
MATCHED-PAIR DESIGN AND RANDOM ASSIGNMENT OF SCHOOLS	17
TARGET STUDENT POPULATION AND CONSENT PROCESS.....	19
SCREENING PROCESS	22
BASELINE EQUIVALENCE	23
CONTAMINATION, CROSSTOVERS, AND STUDENT MOBILITY	29
SAMPLE AT EACH PHASE OF THE STUDY.....	30
MEASURES	32
DATA ANALYSES.....	37
CHAPTER 3: IMPLEMENTING <i>NUMBER ROCKETS</i>	43
DESCRIPTION OF <i>NUMBER ROCKETS</i>	43
TUTOR TRAINING	46
FIDELITY OF IMPLEMENTATION	51
COST OF IMPLEMENTATION.....	58
CHAPTER 4: ESTIMATING THE IMPACT OF <i>NUMBER ROCKETS</i> ON STUDENT ACHIEVEMENT	59
MAINTENANCE OF BASELINE EQUIVALENCE	59
CONFIRMATORY RESEARCH QUESTION FINDINGS	60
SENSITIVITY ANALYSES.....	61

CHAPTER 5: EXPLORATORY ANALYSES FINDINGS	63
EXPLORATORY ANALYSIS 1: EFFECT OF <i>NUMBER ROCKETS</i> BASED ON BASELINE MATHEMATICS PROFICIENCY FOR STUDENTS PARTICIPATING IN <i>NUMBER ROCKETS</i>	63
EXPLORATORY ANALYSIS 2: EFFECT ON LETTER- AND WORD-READING PROFICIENCY FOR STUDENTS PARTICIPATING IN <i>NUMBER ROCKETS</i>	65
EXPLORATORY ANALYSIS 3: RELATIONSHIP BETWEEN PROGRAM IMPACTS AND AVERAGE NUMBER OF DELIVERED LESSONS.....	66
CHAPTER 6: SUMMARY OF KEY FINDINGS AND STUDY LIMITATIONS	68
EFFECT OF <i>NUMBER ROCKETS</i> ON MATHEMATICS ACHIEVEMENT	68
EFFECT OF <i>NUMBER ROCKETS</i> BY BASELINE MATHEMATICS PROFICIENCY OF AT-RISK STUDENTS.....	69
EFFECT OF PARTICIPATION IN <i>NUMBER ROCKETS</i> ON READING ACHIEVEMENT.....	70
EFFECT OF <i>NUMBER ROCKETS</i> BY THE SCHOOL-AVERAGE NUMBER OF TUTORING SESSIONS	70
IMPLEMENTATION OF <i>NUMBER ROCKETS</i> IN A REAL-WORLD CONTEXT	70
STUDY LIMITATIONS AND FUTURE RESEARCH.....	71
APPENDIX A: STUDY TIMELINE.....	A-1
APPENDIX B: POWER ANALYSIS ASSUMPTIONS	B-1
MINIMUM DETECTABLE EFFECT SIZE.....	B-1
ASSUMPTIONS MADE FOR POWER ANALYSIS	B-1
BASIS FOR THE ASSUMPTIONS.....	B-2
TARGET SAMPLE SIZE	B-3
APPENDIX C: PARENT CONSENT FORM	C-1
APPENDIX D: SCREENER SUBTEST DETAILS AND DESCRIPTIVE STATISTICS.....	D-1
APPENDIX E: STUDENT MOBILITY.....	E-1
APPENDIX F: FIDELITY MEASURES	F-1
APPENDIX G: MODELS USED FOR CONFIRMATORY, EXPLORATORY, AND SENSITIVITY ANALYSES	G-1
MULTIPLE IMPUTATION.....	G-1
CONFIRMATORY ANALYSIS	G-1
SENSITIVITY ANALYSES.....	G-3
EXPLORATORY IMPACT ANALYSES	G-4
APPENDIX H: LESSONS.....	H-1
APPENDIX I: COMPLETE SAMPLE LESSON TOPIC 6, DAY 1	I-1
APPENDIX J: DETAILS OF TUTOR TRAINING	J-1
INITIAL TRAINING	J-1
FOLLOW-UP TRAININGS.....	J-4
APPENDIX K: TUTOR BACKGROUND SURVEY.....	K-1
APPENDIX L: DETAILS OF FIDELITY CODER TRAINING	L-1
COMPONENT 1: ORIENTATION TO THE PROGRAM STRUCTURE, ELEMENTS, AND MATERIALS	L-1
COMPONENT 2: INFORMATION ABOUT THE FLASHCARD PROCEDURE	L-2

COMPONENT 3: PRACTICE CODING SAMPLE LESSONS	L-2
COMPONENT 4: LOGISTICS OF COMPLETING LESSON FIDELITY CHECKLISTS OF THE ASSIGNED LESSONS	L-3
APPENDIX M: COMPLETE MULTILEVEL MODEL RESULTS FOR CHAPTER 4 (CONFIRMATORY AND SENSITIVITY) AND CHAPTER 5 (EXPLORATORY AND SENSITIVITY) ANALYSES	M-1
ANALYSES REPORTED IN CHAPTER 4	M-1
ANALYSES REPORTED IN CHAPTER 5	M-5
REFERENCES	REF-1

Tables

Table 1-1. Key differences between the Fuchs et al. (2005) study and the current study	7
Table 2-1. Recruiting summary data across all districts	17
Table 2-2. Participating school sample across all districts.....	19
Table 2-3. Parent consent form return rates for students eligible to receive consent forms, by study condition and school district	21
Table 2-4. Baseline equivalence of student demographics for all schools randomly assigned, for all grades combined and for grade 1, for all 78 schools initially assigned, and for the 76 remaining after attrition.....	25
Table 2-5. Baseline equivalence of screener scores for all screened students and students identified as at risk, by condition and across all districts	27
Table 2-6 Baseline demographic characteristics for all screened students and students identified as at risk, by condition and across all districts	28
Table 2-7. Demographic characteristics and mean screener composite scores for students with TEMA–3 scores and for students missing TEMA–3 scores	42
Table 3-1. Tutor training activities and attrition	50
Table 3-2. Fidelity of lesson implementation by district and implementation-period	54
Table 3-3. Average number of lessons delivered per tutoring group.....	55
Table 3-4. Percentage of tutoring groups that completed intervention Topics 11 through 17	56
Table 3-5. Percentage of reported <i>Number Rockets</i> lessons in which a specified classroom activity was missed	57
Table 4-1. Demographic characteristics and mean screener composite score for students with TEMA–3 scores, by assigned condition	60
Table 4-2. Impact of <i>Number Rockets</i> on mathematics achievement of grade 1 students as measured by the TEMA–3, by assigned condition	61
Table 4-3. Summary of results for the six sensitivity analyses conducted on the impact of <i>Number Rockets</i> on mathematics achievement of grade 1 students, as measured by the TEMA–3, by assigned condition	62
Table A-1. Dates of study phases by district.....	A-1
Table B-1. Power analysis (minimum detectable effect size for minimum of 70 schools).....	B-3
Table D-1. Screener for current study	D-1
Table D-2. Descriptive statistics for the six screener subtests and the screener composite score	D-2
Table E-1. Mobility for students in the analytic sample	E-1
Table F-1. Example of aggregated instructional log data	F-2

Table H-1. <i>Number Rockets</i> , required and additional lessons by topic and day	H-1
Table K-1. Characteristics of mathematics tutors who completed the tutor background survey (<i>n</i> = 75), across all districts	K-3
Table M-1. Confirmatory impact analysis.....	M-1
Table M-2. Sensitivity analysis 1: excluding 26 schools affected by natural disaster.....	M-2
Table M-3. Sensitivity analysis 2: without matched pairs	M-2
Table M-4. Sensitivity analysis 3: without baseline covariate (screener).....	M-3
Table M-5. Sensitivity analysis 4: using cases with complete Test of Early Mathematics Ability–Third Edition (Ginsburg and Baroody 2003) scores only	M-3
Table M-6. Sensitivity analysis 5: excluding students assigned to tutoring groups with students who were not part of the at-risk analytic sample	M-4
Table M-7. Sensitivity analysis 6: excluding school pairs with tutoring groups that included students who were not part of the at-risk analytic sample	M-4
Table M-8. Exploratory 1: differential impact based on baseline mathematics proficiency.....	M-5
Table M-9. Exploratory 2: effect on letter- and word-reading proficiency for students participating in <i>Number Rockets</i> , for cases with complete Woodcock Johnson–Third Edition Letter/Word (Woodcock, McGrew, and Mather 2001) subtest-reading scores only.....	M-5
Table M-10. Exploratory 3: Relationship between implementation level and school-pair level impact of <i>Number Rockets</i>	M-6
Table M-11. Exploratory 1 sensitivity analysis: effect of <i>Number Rockets</i> for lowest third of students at-risk for mathematics difficulties.....	M-6
Table M-12. Exploratory 1 sensitivity analysis: effect for middle third of students at-risk for mathematics difficulties	M-7
Table M-13. Exploratory 1 sensitivity analysis: effect for highest third of students at-risk for mathematics difficulties	M-7
Table M-14. Exploratory 1 sensitivity analysis: using lowest third as the reference group	M-8
Table M-15. Exploratory 1 sensitivity analysis: using highest third as the reference group	M-9

Figures

Figure 2-1. Study timeline	14
Figure 2-2. Schools and grade 1 students in the sample for each phase of the study	31
Figure 3-1. Excerpt from Topic 6, Day 1 lesson <i>Number Rockets</i> script: tutor introduces place value	45
Figure 3-2. Excerpt from Topic 6, Day 1 lesson <i>Number Rockets</i> script: tutor models place value	46
Figure F-1. Sample lesson fidelity checklist	F-1
Figure F-2. Classroom instruction checklist	F-3
Figure I-1. Excerpt from <i>Number Rockets</i> tutoring script	I-1
Figure I-2. Review sheet #5	I-1
Figure I-3. Excerpt from Topic 6, Day 1 lesson <i>Number Rockets</i> script: tutor introduces place value	I-2
Figure I-4. Excerpt from Topic 6, Day 1 lesson <i>Number Rockets</i> script: tutor demonstrates place value	I-3
Figure I-5. Excerpt from Topic 6, Day 1 lesson <i>Number Rockets</i> script: tutor introduces Base-10 blocks ...	I-3
Figure I-6. Excerpt from Topic 6, Day 1 lesson: tutor represents numbers, points awarded	I-4
Figure I-7. Topic 6, Day 1 lesson tutoring sheet 1	I-5

Executive summary

The 2004 reauthorization of the Individuals with Disabilities Education Act (IDEA) approved schools' use of alternative methods for determining student eligibility for special education services. IDEA encourages schools to intervene as soon as there is a valid indication that a student might experience academic difficulties, rather than after performance falls well below grade-level. The Response to Intervention (RtI) framework is an approach for providing instructional support to students at risk for these difficulties.

Underpinning RtI is the concept of intensive early intervention for at-risk students to prevent subsequent academic failure (Glover and Diperna 2007). RtI models typically have three tiers of increasing intensity of instruction (Gersten et al. 2008; Gersten et al. 2009). Tier 1 involves research-based core instruction delivered with high fidelity in the classroom by the classroom teacher and universal screening of all students to determine who should receive additional instructional support. Tier 2 involves focused/intensive instruction, often in small groups, for children at risk for failing in the Tier 1 setting. Tier 3 involves even more intensive instruction for students not responding to the Tier 2 interventions and often comprises individual tutoring, referral to a school psychologist, or both.

Despite increasing interest, there is little research on the effectiveness of recommended best practices in RtI (Gersten et al. 2008; Gersten et al. 2009). Recent large-scale studies have begun to compare the effectiveness of different core Tier 1 curricula (see Agodini et al. 2009), but the evidence for Tier 2 interventions is weaker. Evidence for early mathematics interventions is particularly lacking (Gersten et al. 2009), making it difficult for state and local education agencies seeking to implement RtI models to meet the recommendations for the use of evidence-based practices (Glover and Diperna 2007; Vaughn and Fuchs 2003). Several studies have found that early mathematics achievement is a strong predictor of later mathematics achievement (Duncan et al. 2007; Morgan, Farkas, and Wu 2009), and others have asserted that early intervention is important for improving outcomes for students at risk for mathematics difficulties in the early primary grades (K–3; National Mathematics Advisory Panel 2008).

A recent literature review of grades K–3 mathematics interventions suitable for use in Tier 2 revealed just nine relevant studies (Newman-Gonchar, Clarke, and Gersten 2009), with just one that was a rigorous evaluation of an intervention, and that used a randomized controlled trial (RCT) design (Fuchs et al. 2005). The Fuchs et al. (2005) study examined the impact of *Number Rockets*, a small-group tutoring intervention for grade 1 students at risk for mathematics difficulties, and found statistically significant positive effects on several measures of mathematics proficiency. But that study was an efficacy trial (one implemented under ideal conditions), involved considerable monitoring and support for experienced tutors, and was conducted in a single district.

This study builds on the Fuchs et al. (2005) study and is the first large-scale effectiveness trial (one intended to approximate real-world implementation) of *Number Rockets*. While the Fuchs et al. study was conducted in 10 schools in one district, the current study examined the impact of *Number Rockets* in 76 schools across four districts in four states. While the Fuchs et al. study used tutors experienced with at-risk students, the current study employed tutors with a range of experience who were selected from the local community. While the Fuchs et al. study provided tutors with substantial monitoring and support, the current study provided professional development and a support program similar to those provided by publishers of curriculum products (Agodini et al. 2009). Finally, the district in the Fuchs et al. study used just one curriculum; each of the four urban districts in the current study used a different one, which may have provided a more heterogeneous instructional context.²

The current study addresses the following confirmatory research question:

- Do grade 1 students at risk in mathematics who participate in *Number Rockets* perform better than at-risk control students on the Test of Early Mathematics Ability—Third Edition (TEMA–3; Ginsburg and Baroody 2003)?

The study also investigated three exploratory research questions:

- Does *Number Rockets* have a differential impact on grade 1 students at risk in mathematics, based on baseline mathematics proficiency?
- Do grade 1 students who participate in *Number Rockets* score differently than control students on the Woodcock-Johnson—Third Edition Letter/Word (WJ–III Letter/Word; Woodcock, McGrew, and Mather 2001) subtest?
- Do the impacts of *Number Rockets* vary significantly depending on the average number of lessons delivered within a school?

Thus, the first exploratory research question examined whether the effect of *Number Rockets* depended on student baseline mathematics proficiency. The second examined whether intervention students, who missed some regular classroom instruction when attending *Number Rockets* tutoring sessions, scored differently on a measure of word reading skill than control students, who did not miss regular classroom instruction. The third examined whether the school-level intervention effect on student TEMA–3 performance varied by the average number of tutoring sessions at each intervention school.

² Each participating district used a different core mathematics curriculum: enVision Math™, Houghton Mifflin Math™, Math Investigations™, or Scott Foresman-Addison Wesley.™

Description of the program

Number Rockets is a scripted tutoring Tier 2 intervention program for grade 1 students identified as at risk through universal classroom screening. Its goal is to build students' conceptual understanding by beginning with concrete tasks and transitioning to representational activities and later to more abstract tasks. A tutor delivers the program to groups of two or three students who meet outside the classroom during the regular school day, though not during regular mathematics instruction. *Number Rockets* is thus a tradeoff for participating students; they can benefit from the additional mathematics instruction but lose instruction in another subject. *Number Rockets* has 63 lessons covering 17 topics, including sequencing numbers, skip counting, and place value. Each session lasts about 40 minutes (a half-hour content lesson and 10 minutes of mathematics fact practice using flashcards).

Study design, methodology, and implementation

This RCT was implemented in 76 schools in four urban districts across four of the five Regional Educational Laboratory Southwest states. *Number Rockets* is implemented at the school level; so schools were the unit of random assignment. They were matched within district on a composite score calculated from mean school achievement scores and the percentage of students receiving free or reduced-price lunch. One member of each school pair was then randomly assigned to the intervention condition; the other, to the control condition.

The target student population was grade 1 students at risk for mathematics difficulties who received mathematics instruction in English in a regular education classroom. Parent consent forms were distributed to eligible students, with a consent rate of 62.0 percent. Consent rates were higher for intervention schools (70.6 percent) than for control schools (52.5 percent). Because schools were randomly assigned to either the intervention or control condition before consent forms were distributed, the differential consent rates could have resulted from varying effort by school personnel in collecting the forms.

All students with consent (2,719 students: 1,643 intervention, 1,076 control) were screened using a measure composed of six subtests. Three subtests were used in the Fuchs et al. (2005) study and measure grade 1 mathematics skills in solving computation problems, concept/application problems, and brief story problems; the other three were selected from research on valid screening measures in mathematics for grade 1 students (Jordan et al. 2007; Geary 1993; Baker et al. 2006; Clarke et al. 2006) and measure number sense, comparative judgments of numerical magnitude, and working memory. Administration time was about 25 minutes. Students with a screener composite score at

or below the sample's 35th percentile (994 students; 615 intervention, 379 control) were considered at risk and participated in the study.

Analyses were conducted to determine whether there were statistically significant differences in overall demographics of study schools or between baseline mathematics proficiency and demographics of the students. Statistically significant differences between intervention and control schools were observed for race/ethnicity, both for grade 1 students and for all grades combined. In addition, grade 1 enrollment was significantly higher in intervention schools. Additional analyses examined all screened grade 1 students and students identified as at-risk. These analyses found statistically significant differences in race/ethnicity between intervention and control schools for all screened students; however, no statistically significant differences were found between intervention and control students identified as at risk (those in the analytic sample).

At-risk students in intervention schools were assigned to tutoring groups of two or three students by study staff, based on tutor availability and school and classroom schedules. Tutoring groups met three or more times per week for approximately 17 weeks. *Number Rockets* was delivered in addition to regular core mathematics instruction. At-risk students in control schools received regular core mathematics instruction but no additional support (the counterfactual condition). The target minimum number of lessons to be delivered was 45, and, on average, 48.4 lessons were delivered to each tutoring group, resulting in approximately 32 hours of intervention time. At the end of the intervention, the TEMA-3 (a broad measure of student proficiency in mathematics) and the WJ-III Letter/Word subtest (a reading fluency measure) were used to collect posttest data for 90 percent (555 out of 615) of intervention students and 86 percent (326 out of 379) of control students. Students who were not available during the post-testing window due to absence or mobility were not assessed.

The TEMA-3 was selected as the primary outcome measure because it represented a broad measure of mathematics achievement for grade 1, this type of outcome is of high interest to educators and policy-makers. The WJ-III Letter/Word subtest was selected due to its wide-use by researchers, and as a secondary outcome measure its characteristics provided a balance between brevity and sensitivity to reading fluency appropriate for the exploratory research question.

The effect of *Number Rockets* on TEMA-3 performance was estimated by comparing at-risk students in the intervention group with their control group counterparts. The analyses were conducted using hierarchical linear modeling, an approach that accounts statistically for the clustered data in this study (students clustered within schools; intervention schools matched to control schools). An intent-to-treat approach was used to analyze student data based on the study condition to which their school was randomly assigned, regardless of whether a student was treated as intended. Students with

missing posttest data were included in the analyses through multiple imputation, an approach used to address missing data.

Analysis and findings

This study's confirmatory finding was that at-risk grade 1 students participating in *Number Rockets* had significantly higher TEMA-3 scores than at-risk grade 1 students in the control group (effect size = 0.34, $p < .001$). Six sensitivity analyses were conducted and found that the confirmatory impact estimate was robust to the analytic choices examined.

None of the three exploratory analyses found significant effects. The first found that the effect of *Number Rockets* did not depend on student baseline mathematics proficiency, as determined by screener composite score (not statistically significant; effect size=0.08, $p = .564$). The second found that intervention group students (who missed regular classroom instruction while participating in *Number Rockets*) did not score significantly different on the WJ-III Letter/Word subtest than control students who did not participate in *Number Rockets* (effect size = -0.01, $p = .913$). The third found no significant relationship between the average number of *Number Rockets* tutoring sessions delivered to each intervention school and the school-level intervention effect (effect=0.07, $p = .667$). However, given that a greater portion of the variability in sessions delivered to student groups occurred within schools than between schools, this exploratory analysis is not sensitive enough to rule out the existence of a dosage-impact relationship at the school-pair level. Note that the study was not specifically designed or powered for the exploratory research questions.

Conclusions

The main finding of this effectiveness study is that grade 1 students at-risk for difficulties in grade 1 mathematics benefited from participation in the *Number Rockets* intervention. At-risk students in the intervention group showed statistically significant higher performance on the TEMA-3, a broad measure of student proficiency in mathematics, than at-risk students in the control group. This finding was observed in a sample of 994 students from 76 schools in four urban districts across four states. The results of all three exploratory analyses (related to differences in baseline mathematics proficiency, performance on a reading test, and number of tutoring sessions) were not statistically significant.

Limitations and suggestions for future research

Several limitations must be considered.

- First, the counterfactual condition in this study consisted of regular classroom instruction and no added mathematics instruction for at-risk students. It cannot be stated whether the intervention effect was due to additional mathematics instruction time delivered in any manner or to the design of *Number Rockets*.
- Second, requiring parent consent introduced a potential student selection bias after schools were randomly assigned, and differential consent form return rates were observed between the intervention and control schools. While students in the at-risk analytic sample intervention and control groups (the subsample of all screened students upon which this evaluation is based) did not differ statistically on observed demographic characteristics or screener composite scores, whether the observed differential consent form rates influenced the baseline equivalence of the two experimental groups on unobserved characteristics is unknown.
- Third, specific urban districts were recruited for this study, and the students included represented a sample whose parents gave consent for student participation. Because districts and schools volunteered for the study, the districts and schools are not statistically representative of a larger population.
- Fourth, *Number Rockets* is not available in Spanish. In study districts, English-language learner students comprised from 1 percent to 29 percent of students across all grades (National Center for Education Statistics; n.d.).
- Fifth, the current study focused on outcomes at the end of grade 1. This study does not provide evidence on the persistence of the benefits of *Number Rockets*, and it is unknown whether students who benefited in grade 1 would be better prepared for success in mathematics at the beginning of grade 2 or beyond.
- Finally, tutors were instructed not to communicate information about individual student performance to classroom teachers, a constraint imposed to prevent contamination of *Number Rockets* strategies into the classroom. This rule might be relaxed in a real-world implementation of a Tier 2 intervention, allowing classroom teachers to have regular communication with tutors about how students from their classrooms are performing.

The increasing interest in RtI, lack of evidence supporting mathematics RtI Tier 2 interventions relevant for grades K–3, and the statistically significant positive effect of *Number Rockets* on the mathematics achievement of at-risk grade 1 students in this study suggest that follow-up studies are warranted. It would be important to compare *Number Rockets* with a counterfactual condition that controlled for added mathematics instruction, either through adding time with the existing mathematics curriculum or using

another Tier 2 intervention to supplement regular instruction. Also, effectiveness studies of *Number Rockets* using representative samples of Southwest Region districts and schools, or the nation as a whole, would allow results to be generalized to a larger population. Replicating this study in other regions (without necessarily attempting to sample districts and schools) could provide evidence of generalizability as well. Follow-up studies could also be conducted to examine the long-term impacts of *Number Rockets* in later grades. Also, given the increasing number of Spanish-speaking students in many schools across the country (National Center for Education Statistics 2004), a Spanish version of *Number Rockets* and subsequent efficacy and effectiveness research would appear to be of value.

Further studies could be undertaken to examine the tradeoffs for the level of tutor professional development (for example, evaluating whether a shorter tutor training regimen would be as effective as the professional development in the current study). Future studies could also examine the tradeoffs for the level of tutoring provided (for example, evaluating whether a 20–30 minute intervention would be as effective as the 40-minute intervention evaluated in the current study). Other studies could be designed and powered to examine dosage effects based on the number of sessions delivered to tutoring groups, to determine if there is a minimum number of sessions required to achieve the impact observed in this study.

Chapter 1: Introduction and study overview

The 2004 reauthorization of the Individuals with Disabilities Education Act (IDEA) approved schools' use of alternative methods for determining student eligibility for special education services. IDEA encourages schools to intervene as soon as they expect a student might experience academic or behavioral difficulties, rather than after performance falls well below grade level. A Response to Intervention (RtI) framework is an approach for providing instructional support to students at risk for academic difficulties. In this framework, there are typically three tiers of increasing instructional intensity (Gersten et al. 2008; Gersten et al. 2009; Vaughn and Fuchs 2003). Tier 1 involves regular classroom instruction, evidence-based if possible; Tier 2 involves more intensive instruction, often delivered to small groups; and Tier 3 involves even more intensive instruction, typically individualized and possibly one-on-one, for students struggling even with Tier 2 intervention. Both Tiers 2 and 3 supplement regular classroom instruction.

Each Regional Educational Laboratory (REL) Southwest Region state (Arkansas, Louisiana, New Mexico, Oklahoma, Texas) has developed state-level policies or guidance on the use of RtI.³ But as documented in two recent Institute of Education Sciences practice guides (Gersten et al. 2008; Gersten et al. 2009), there is limited research on RtI. The evidence on early mathematics interventions is particularly lacking when compared with reading (Gersten et al. 2009), making it difficult for state and local education agencies to implement RtI models that meet the recommendations for the use of evidence-based practices (Glover and Diperna 2007; Vaughn and Fuchs 2003).

Number Rockets is a supplemental Tier 2 mathematics intervention for grade 1 students considered at risk for difficulties in mathematics. Previous research (Fuchs et al. 2005) examining *Number Rockets* demonstrated statistically significant positive effects on computation and concepts/applications skills for these students. Twenty-one percent of students with consent were designated as at risk based on a two-stage screening process: poor performance on a battery of four mathematics achievement tests and poor performance four weeks later on a brief assessment of computation skills. At the end of grade 1, there was a statistically significant difference in favor of *Number Rockets* students, compared with control students, on four tests measuring computational and concepts/application skills; however, there was not a statistically significant difference on a test of applied problems and two tests of mathematics fact fluency (Fuchs et al. 2005). In that efficacy study—a study in which the intervention is implemented to developer specifications with high fidelity—the *Number Rockets* developers were directly involved in implementation, and the study was conducted in only one district.

³ Arkansas State Department of Education Special Education Unit 2010; Louisiana Department of Education 2009; New Mexico Public Education Department 2009; Oklahoma State Department of Education 2007; Texas Education Agency 2008.

This report builds on that study by examining Number Rockets under conditions more closely resembling the experiences of school districts in their day-to-day instructional environments when implementing interventions (an effectiveness trial). This report represents the first large-scale effectiveness trial of Number Rockets.

Importance of early intervention in mathematics

Recent studies have found that early mathematics achievement is a strong predictor of later mathematics achievement (Duncan et al. 2007; Morgan, Farkas, and Wu 2009). In an analysis of the Early Childhood Longitudinal Study–Kindergarten Cohort (ECLS–K) dataset,⁴ 65 percent of ECLS–K students identified in kindergarten as having mathematics difficulties (those in the lowest 10 percent of the sample) were still in the lowest 10 percent of the study sample four years later (Morgan, Farkas, and Wu 2009). Duncan et al. (2007) conducted a meta-analysis of six studies examining different datasets (including the ECLS–K) and found the correlation between mathematics performance at school entry and later mathematics performance⁵ to be 0.47 on average; early mathematics proficiency was a better predictor of later mathematics achievement than were attention skills or socio-emotional skills.

The strength of the association between early and later mathematics achievement provides a rationale for schools to focus additional instruction resources on students in prekindergarten and the early primary grades (K–3) who are struggling to progress through mathematical concepts deemed important for later mathematics achievement (National Council of Teachers of Mathematics 2006; National Mathematics Advisory Panel 2008; National Research Council 2009). A strong K–12 mathematics education is important for ensuring the future stability of the U.S. economy, national security, and workforce productivity (National Council of Teachers of Mathematics 2009); “an economically competitive society recognizes the importance of mathematics to adult numeracy and financial literacy, and it depends on citizens who are mathematically literate” (National Council of Teachers of Mathematics 2009, p. 1).

⁴ The ECLS–K was designed around a large, nationally representative sample of children. A cohort of students who entered kindergarten in fall 1998 has been followed longitudinally, with data made available through grade 8 as of 2010. The ECLS–K uses vertically-scaled mathematics assessments based on National Assessment of Educational Progress specifications, which provide item response theory scores appropriate for modeling growth over time (Morgan, Farkas, and Wu 2009).

⁵ Across the six studies, school-entry measures were collected from students 4.5 through 5 or 6 years of age. Outcome measures were collected from students ages 8–9 years to ages 13–14 years, depending on the study.

Response to Intervention

RtI is one strategy for schools to address issues of academic performance, including mathematics performance, in the early grades. RtI is not only a process for determining special education eligibility but also a schoolwide model for helping students struggling academically (Clarke, Gersten, and Newman-Gonchar 2010; Vaughn, Linan-Thompson, and Hickman 2003). Underpinning RtI is intensive early intervention to prevent later academic failure (Glover and Diperna 2007). RtI models typically have three tiers of increasing instructional intensity (Gersten et al. 2008; Gersten et al. 2009), though some define the activities performed in each tier differently and may have two or four tiers (Barnes and Harlacher 2008; New Mexico Public Education Department 2009; Tilly 2003). RtI models could be implemented in any content area where universal screening (that is, screening of all students in the school for at-risk status) is possible, but to date have been implemented primarily in reading and mathematics. For this study, the tiers will be discussed in the context of mathematics instruction.

Tier 1 is the mathematics instruction all students receive and consists of research-based (where possible) core instruction delivered in the classroom. Although not all experts agree on the characteristics of Tier 1, most agree that it should include differentiated instruction⁶ for students having difficulties (Gersten et al. 2009). Universal screening of all students, regardless of mathematics proficiency, to determine those students likely to need instruction beyond Tier 1, is a critical feature of RtI.

Tier 2 is typically focused, intensive instruction often delivered in small groups of two or three students meeting two or three times per week. It should be provided in addition to regular whole class instruction (Gersten et al. 2009). Tier 2 often comprises an increased level of targeted instruction in specific mathematics skills (Fuchs et al. 2008). Instruction at this tier can be provided outside the student's regular classroom by such people as the classroom teacher, a classroom aide, an instructional specialist, or a tutor. Student progress is monitored (typically weekly or biweekly) to determine whether the student no longer needs Tier 2, should continue with Tier 2, or advance to Tier 3.

The distinction between Tier 2 and Tier 3 intervention is primarily the intensity (Gersten et al. 2009), represented by the size of instruction group, amount of instruction time, and number and type of school personnel involved. Tier 3 instruction is also provided in addition to regular classroom instruction, often daily and one-on-one (for example, see O'Connor, Harty, and Fulmer 2005). Additional personnel such as school

⁶ Differentiated instruction is an approach that uses strategies such as flexible student groupings designed to facilitate learning that do not require specific amounts of time or assessment before movement into and between such groups takes place. This approach and related strategies often consist of whole-class instruction, where individual or small group modifications occur only as the teacher judges them necessary for a student's need or the teacher is "convinced that modification increases the likelihood that the learner will understand important ideas and use important skills more thoroughly as a result" (Tomlinson 1999, p. 11).

psychologists or special education specialists may be incorporated (Barnes and Harlacher 2008). Student progress is monitored (typically weekly or biweekly) to determine if students should continue with Tier 3, move back to Tier 2, or be formally recommended for special education services.

Response to Intervention Tier 2 mathematics intervention research

Recent large-scale studies have compared the effectiveness of core Tier 1 mathematics curricula (Agodini et al. 2009); however, evidence on Tier 2 interventions is lacking. Researchers have emphasized the importance of students receiving evidence-based classroom instruction (Tier 1) in conjunction with validated Tier 2 interventions (Fuchs et al. 2008). Identifying effective Tier 2 interventions that can mitigate the potential need for a one-on-one Tier 3 intervention or special education services can benefit both students and schools. Being designated for special education “can be problematic when it stigmatizes students, separates them from their peers, results in lower academic expectations, generates undesirable educational outcomes . . . [in addition to] the immense direct and lost opportunity costs . . . Students who become categorized into one of the learning disabilities categories that makes them eligible for special education rarely shed that label through the course of their education” (Hruz 2002, p. 26). From a cost perspective alone, preventing inaccurate learning disorder identification can save schools substantial amounts of resources.

REL Southwest identified the need for a large-scale regional evaluation of Tier 2 mathematics intervention. When the current study was designed, a systematic review of research published from 1990–2007 (Newman-Gonchar, Clarke, and Gersten 2009) had identified only two studies (Fuchs et al. 2005; Bryant et al. 2008) evaluating impacts of Tier 2 mathematics interventions for grade 1 students.⁷ Both interventions studied represented candidates for further investigation in a large-scale evaluation.⁸ Bryant et al. (2008) included 126 grade 1 and 140 grade 2 students from one school; a statistically significant impact was not found for the mathematics achievement of grade 1 students ($b = 0.04$, n.s.), but was found for grade 2 students ($b = 0.19$, $p = .018$).⁹ Because Bryant et

⁷ Grade 1 is the first time all students receive formal Tier 1 core mathematics instruction. Pianta et al. (2008) found that most growth in applied problem solving in the elementary grades happens before grade 3.

⁸ An intervention was a candidate for large-scale evaluation if it had an evidence base for grade 1 outcomes that included well-designed randomized controlled trials or strong quasi-experimental studies, such as regression discontinuity designs.

⁹ The Bryant et al. (2008) study used a regression discontinuity design and reported standardized betas (b). In a linear regression model with one predictor, the unstandardized beta coefficient (β) is the correlation between the predictor (independent) variable and the outcome (dependent) variable. A standardized beta is simply β converted to a mean of 0 and variance of 1. No p -value was reported for grade 1, precluding conversion to an effect size. The alpha used was not specified; however, the result was not statistically significant.

al. (2008) used the less rigorous regression discontinuity design and was conducted in one school, the intervention examined was considered less suitable for large-scale evaluation than that in the Fuchs et al. (2005) study and identified in the review of previous research.

The Fuchs et al. (2005) study was a randomized controlled trial (RCT) examining the impact of a Tier 2 intervention in improving grade 1 mathematics achievement. Ten public schools in one district participated, and 667 students were screened to determine at-risk status. Approximately 21 percent ($n = 139$) of students screened were identified as at risk and randomly assigned to the intervention or control group. The intervention group received approximately 17 weeks of supplemental small-group tutoring, and missed the classroom instruction that took place during that time. The Fuchs et al. study did not specify what classroom instruction—other than mathematics—was missed by intervention students. Tutors were instructed that students should not miss regular classroom mathematics instruction. Still, intervention students missed an average of 10.56 minutes of regular mathematics instruction in total during the intervention. Students in the control (or counterfactual) condition remained in classrooms during the intervention time and did not receive mathematics instruction beyond the regular classroom. The intervention and control students were assessed at the end of the intervention period on seven measures of mathematics skills. Statistically significant results favoring the intervention group were identified for four of the seven; effect sizes¹⁰ for the seven measures ranged from 0.11 to 0.70.¹¹

The positive impacts observed in the Fuchs et al. (2005) study indicate that small group tutoring is an ideal intervention for the next step—evaluation on a larger scale under conditions more closely resembling the experiences of school districts in their day-to-day instructional environment when implementing interventions. After the 2005 study was published, the intervention was titled *Number Rockets* by the developer (Lynn Fuchs, Professor and Nicholas Hobbs, chair in special education and human development, Vanderbilt University—personal communication, August 6, 2009).

¹⁰ An *effect size* is a standardized measure of the strength of an outcome. There are several ways to calculate effect sizes, but the primary approach used in this report is Hedges' *g*, in which the difference in group means (that is, treatment mean minus the control mean) is divided by the pooled standard deviation. This effect size provides the difference in standard deviation units. (For example, an effect size of 0.50 implies that the difference in group means is one-half of the pooled standard deviation.). Effect sizes in Fuchs et al. (2005) are reported as Cohen's *d*, which has been shown to be "upwardly biased when based on small sample sizes" (p. 48, Lipsey and Wilson, 2001). Hedges's *g* is closely related, but includes a small correction for this bias, and the two measures can be interpreted in the same manner.

¹¹ Effect sizes reported in the Fuchs et al. (2005) study were calculated as change scores, subtracting posttest means from pretest means and correcting for the correlation between the two. Measures are listed in order of effect size, and the four with statistically significant effect sizes are listed first: Story Problems (effect size = 0.70); First-Grade Concepts/Applications (0.67); Woodcock-Johnson Third Edition (WJ-III)—Calculation (0.57); Curriculum-Based Measurement—Computation (0.40); Addition Fact Fluency (0.40); Subtraction Fact Fluency (0.14); and WJ-III—Applied Problems (0.11).

Description of the program

Number Rockets, a scripted program available only in English, consists of 63 lessons across 17 mathematics topics (3–6 lessons on each topic). Each lesson lasts about 40 minutes (30 minutes of instruction and 10 minutes of fact practice). The intervention is delivered to groups of two or three students, similar to the intensive systematic small group instruction suggested for use in Tier 2 instruction (Gersten et al. 2009). *Number Rockets* is a supplemental intervention, provided in addition to core classroom mathematics instruction and not intended to replace it. So schools implementing *Number Rockets* should not remove students from regular mathematics instruction to participate in tutoring sessions. Students can, however, be pulled out of regular instruction in other subjects to attend *Number Rockets* lessons.¹²

As described by the developers (Fuchs et al. 2005) *Number Rockets* was designed to emphasize the development of number sense¹³ and to build conceptual understanding and procedural fluency¹⁴ with whole numbers, both critical to mathematics progress after grade 2 (National Research Council 2001, 2009; National Mathematics Advisory Panel 2008). The intervention stresses both concepts and operations involving whole numbers including number sense, computation, and place value. Rather than span the entire grade 1 mathematics curriculum, it has a narrow focus so that critical topics can be taught in depth (Lynn Fuchs, Professor and Nicholas Hobbs chair in special education and human development, Vanderbilt University—personal communication, August 6, 2009). This is an approach widely advocated by experts (National Council of Teachers of Mathematics 2006; National Mathematics Advisory Panel 2008). See chapter 3 for a detailed description.

Current study

The current study, a multidistrict effectiveness RCT conducted in the REL Southwest Region, examines the impact of *Number Rockets* on the mathematics achievement of at-risk grade 1 students. Students were clustered within schools, and schools were the unit of random assignment.

¹² Chapter 3 presents data describing the degree to which schools in the present study followed this recommendation.

¹³ The Kalchman, Moss, and Case (2001, p. 2) study defines the characteristics of good number sense to include “(a) fluency in estimating and judging magnitude, (b) ability to recognize unreasonable results, (c) flexibility when mentally computing, and (d) ability to move among different representations and to use the most appropriate representation.” However, there is not perfect agreement on the definition of number sense; Berch (2005) lists 30 elements that various researchers have claimed represent the construct.

¹⁴ Procedural fluency is the “skill in carrying out procedures flexibly, accurately, efficiently, and appropriately” (National Research Council 2001, p. 116).

The current study differs from the Fuchs et al. (2005) study in several ways. First, the current study is an effectiveness trial examining implementation under conditions that more closely resemble what school districts experience in their day-to-day instructional environment when implementing interventions; the Fuchs et al. study was an efficacy trial examining implementation in ideal conditions. This impacted the tutor training and coaching, as well as the fidelity of implementation observed. (See chapter 3 for more information.) In addition, the current study was implemented in four districts across four states; the Fuchs et al. study took place in a single district. Both studies were conducted entirely in public schools. An overview of key differences between the studies is in table 1-1; additional details of the current study's design are in chapter 2.

Table 1-1. Key differences between the Fuchs et al. (2005) study and the current study

<i>Characteristic</i>		<i>Fuchs et al. (2005)</i>	<i>Current study</i>
Study	Type	Efficacy trial	Effectiveness trial
	Design	Student-level random assignment within classroom	School-level random assignment, schools paired within district
Consent	Type	Active parent consent	Active parent consent
	Rate	89 percent consent granted ^a	70.6 percent consent granted for intervention students 52.5 percent consent granted for control students
Sample	Districts	One district	Four districts across four states
	Schools	10 urban public elementary schools	76 urban public elementary schools
		6 Title I schools, 4 non-Title I schools	73 Title 1 schools, 3 non-Title 1 schools
	Students	667 screened 139 identified as at-risk 70 received the intervention 69 served as controls	2,719 screened 994 identified as at-risk 615 received the intervention 379 served as controls
Screening	Procedure	Two stages: (1) A 15-minute screener comprised of four mathematics tests (2) Response to classroom instruction measured by limited progress on weekly CBM ^b measures after 4 weeks Students rank-ordered by factor score	One stage: A 25-minute screener comprised of six mathematics tests Students rank-ordered by composite score
	At-risk rate	Lowest 21 percent of students screened	Lowest 35 percent of students screened
Teacher involvement		Trained the regular classroom teacher to administer weekly CBM measures to whole class	Only teachers in intervention schools knew students' at-risk status No progress monitoring data were

<i>Characteristic</i>		<i>Fuchs et al. (2005)</i>	<i>Current study</i>
Tutors		Teachers provided with student progress monitoring reports and classroom instructional strategies by research team every 2 weeks	collected Teachers received no information about students' progress
	Total number	12	86
	Qualifications	10 Master's-level graduate students 1 Ph.D. researcher 1 experienced tutor	100 percent with a bachelor's degree (at minimum) Wide range in teaching experience: 6 months to 38 years Locally recruited from retired teachers and substitute teacher pool
	Training and coaching	One-day training followed by additional practice and two follow-up sessions prior to tutoring ^c Weekly coaching sessions throughout the intervention ^c	One day (8 hour) training Two 2-hour follow-up trainings with question and sessions Questions submitted and answered via email or telephone, as received
Delivery of mathematics fact practice ^d		Mathematics fact practice delivered via a computer program titled Math Flash	Mathematics fact practice delivered via paper flash cards
Number of lessons		48 lessons delivered	45 lessons targeted for delivery
Primary outcome measure(s)		(1) First Grade Concepts/Applications ^e (2) CBM—Computation ^e (3) Addition Fact Fluency ^f (4) Subtraction Fact Fluency ^f (5) Woodcock-Johnson Third Edition (WJ-III) —Calculation ^g (6) WJ-III—Applied problems ^g (7) Story Problems ^h	Test of Early Mathematics Ability—Third Edition (TEMA-3; Ginsburg and Baroody 2003)

Note: CBM is Curriculum Based Measurement

a. Did not report consent rate by experimental condition; parents provided consent prior to random assignment.

b. CBM—Computation is a 1-page set of 25 grade 1 computation items group-administered to all students weekly in the Fuchs et al. (2005) study for purposes of progress monitoring.

c. Training was provided for one day, followed by additional practice over two weeks; a second training session on how to deliver mathematics fact practice; a final review session prior to tutoring; and weekly coaching meetings. The number of hours involved in the training sessions, the additional practice, and the weekly coaching meetings was not specified.

d. Because of the lack of available computers in study schools, the *Number Rockets* developers adapted Math Flash to a parallel paper format.

e. Fuchs, Hamlett, and Fuchs (1990).

f. Fuchs, Hamlett, and Powell (2003).

g. Woodcock, McGrew, and Mather (2001).

h. Jordan and Hanich (2000).

Source: Fuchs et al. 2005; authors' analysis of data collected October 2007–May 2009.

This study was designed and powered to address the confirmatory research question,¹⁵ and the determination of the effectiveness of *Number Rockets* is based solely on the findings for this question:

- Do grade 1 students at risk in mathematics who participate in *Number Rockets* perform better than at-risk control students on the Test of Early Mathematics Ability—Third Edition (TEMA–3; Ginsburg and Baroody 2003)?¹⁶

Three exploratory research questions were also addressed:

- Does *Number Rockets* have a differential impact on grade 1 students at risk in mathematics, based on baseline mathematics proficiency?
- Do grade 1 students who participate in *Number Rockets* score differently than control students on the Woodcock-Johnson—Third Edition Letter/Word (WJ–III Letter/Word; Woodcock, McGrew, and Mather 2001) subtest?
- Do the impacts of the Number Rockets program vary significantly depending on the average number of lessons delivered?

Structure of the report

Chapter 2 details the study design and methodology, describing the study participants, data collection measures, and data analysis methods. Chapter 3 describes how *Number Rockets* was implemented and addresses the resulting fidelity of implementation. Chapter 4 discusses the maintenance of baseline equivalence for the analytic sample, empirical findings that address the confirmatory research question, and the results of the corresponding sensitivity analyses. Chapter 5 presents the findings of exploratory analysis 1, the associated sensitivity analysis, and the findings of exploratory analyses 2 and 3. Chapter 6 summarizes key findings, describes the study limitations, and suggests directions for future research.

¹⁵ A confirmatory research question is defined as a “precisely stated research hypothesis” which establishes how “the causal effects of an intervention . . . [should] be rigorously estimated using an experimental design” (Burghardt et al. 2009, p.1).

¹⁶ The TEMA–3 (Ginsburg and Baroody 2003) is an individually administered test of mathematics achievement aligned with contemporary thinking about what is essential to teach in mathematics. It has received positive reviews (Crehan 2005; Monsaas 2005) and is widely used (Hojnoski, Silberglitt, and Floyd 2009; Methe, Hintze, and Floyd 2008; Baroody, Li, and Lai 2008).

Chapter 2: Study design and methodology

This chapter begins by explaining the study design, study timeline, power analysis, and sample size. Next, it describes the target district population, the recruitment process, and the matched-pairs design and assignment of schools to conditions. The chapter then addresses the target student population, the consent process, and the screening processes used to identify the student sample. Baseline equivalence is discussed next; followed by contamination, crossovers, and student mobility; and study sample by each phase of the study. The chapter concludes by discussing the data collection measures and data analysis methods.

Study design overview

Seventy-six schools in four urban school districts in four states in the REL Southwest Region participated in this study. Schools were the unit of assignment because Response to Intervention (RtI) and a Tier 2 intervention such as *Number Rockets* would typically be implemented at the school level. Schools were matched within district on a composite score calculated from a mean school mathematics achievement score and the percentage of students receiving free or reduced-price lunch (FRPL). One school in each pair was then randomly assigned to the intervention condition and the other to the control condition.

At each school, students whose parents signed a consent form were screened using an individually administered screener. A simple composite was computed based on the average z -scores¹⁷ across the screener's six subtests. Next, a cutscore for identifying students at risk for mathematics difficulties that corresponded to the 35th percentile of sample students was determined; the 35th percentile cutscore was used because it is consistent with others in the literature. The selection of a relative percentile rank cutscore on the subtest composite was directed primarily by the internal needs of the study. The percentile cutscore is not directly related to specific student skills or abilities related to grade-level mathematics performance. Therefore, some students identified as at-risk for the purposes of this study may not be considered to be at-risk using external criteria, such as a lack of a specific grade 1 mathematics skill like understanding place value. Hanich et al. (2001) assessed grade 2 student reading and mathematics achievement with the Woodcock-Johnson Tests of Educational Achievement (Woodcock and Johnson 1990). Students with a composite at or below the 35th percentile in the study sample were classified as having mathematics difficulties, and students with a reading composite at or below the 35th percentile in the study sample were classified as having reading

¹⁷ A z -score is a standardized measure of the distance between a single data point and the sample mean. A z -score is calculated by subtracting the single value from the sample mean and then dividing that difference by the standard deviation of the sample.

difficulties. In Jordan, Kaplan, and Hanich (2002), grade 2 students were screened using the Woodcock–Johnson Psycho-Educational Battery—Revised, Form A (Woodcock and Johnson 1990). Students were given the mathematics composite subtests and the reading composite subtests. Students with a mathematics composite at or below the 35th percentile in the study sample were classified as having mathematics difficulties and students with a reading composite or a Letter/Word identification subtest score (part of the reading composite) at or below the 35th percentile in the study sample were classified as having reading difficulties. In follow-up study, Jordan, Hanich, and Kaplan (2003) assessed grade 3 students who had previously been screened using the reading and mathematics composites from the Woodcock-Johnson Psycho-Educational Battery—Revised, Form A (Woodcock and Johnson 1990). Students who scored at or below the 35th percentile in the study sample were considered at risk for analysis in the 2003 study.

Because the Fuchs et al. (2005) study used the 21st percentile as the cutscore, the study team considered doing the same thing. But doing so would have reduced the sample size of at-risk students in some schools to zero, due to natural variability in the distribution of student mathematics ability in each school. There was also a lower overall consent rate for the current study (62 percent) than for the Fuchs et al. study (89 percent), meaning that fewer students in the current study were available for screening for at-risk status. So using the 21st percentile instead of the 35th percentile would have caused some schools to be eliminated, reducing the number of schools below that required for the targeted statistical power.

In the present study, at-risk students in intervention schools were assigned to tutoring groups based on student class schedule and group size. Two or three students, the group size for which *Number Rockets* was designed, was also used in the Fuchs et al. (2005) study. Small group tutoring allows students more practice and immediate feedback and correction (Bryant et al. 2008) and is thus recommended in intervention implementation for students in the primary grades (Baker, Gersten, and Lee 2002; Butler et al. 2003; Vaughn, Moody, and Shumm 1998).

Students from the same classroom were assigned to the same tutoring group; if there were more than three at-risk students in one classroom, groups of two students were created as needed. As a condition of participation, intervention schools were promised that a minimum of nine students would receive *Number Rockets*; to satisfy this commitment, some tutoring groups included students who did not meet the at-risk criteria. These students were not included in analyses and were not posttested. Intervention students participated in the *Number Rockets* program three or more times per week for approximately 17 weeks. In the Fuchs et al. (2005) study, 48 lessons were delivered to each tutoring group. Because the current study was designed as an effectiveness trial, the study team established a lower target of 45. This target allowed a two-week buffer for district holidays between the December 1 start date for implementation and the posttest dates established with the districts. It also allowed for

some scheduling flexibility so that the study team was not at the same time attempting to collect posttests in all four districts.

Number Rockets was designed to supplement core mathematics curricula, not replace regular classroom instruction. The intervention does, however, cover similar concepts and skills incorporated in most grade 1 core mathematics curricula. At-risk students in control schools received regular core mathematics instruction but no additional support beyond what a grade 1 teacher might provide to students experiencing difficulties (Tier 1 instruction). The lack of other supplemental grade 1 mathematics programs supported by a rigorous evidence base precluded the selection of a different counterfactual in this study.¹⁸ As a condition of enrollment in the study, schools agreed that at-risk students participating in *Number Rockets* would not miss regular mathematics instruction. However, absence during mathematics instruction occurred to some degree. (See chapter 3 for more information.) The Test of Early Mathematics Ability–Third Edition (TEMA–3; Ginsburg and Baroody 2003) was individually administered to at-risk students in the intervention and control groups at the end of the intervention period.

This study used an intent-to-treat (ITT) approach, estimating the impact of *Number Rockets* for students offered to participate in (as opposed to actually receiving) the intervention. In particular, the impact analysis compares the mathematics achievement of at-risk students randomly assigned to receive regular classroom mathematics instruction and *Number Rockets* (the intervention group) with at-risk students randomly assigned to receive only regular classroom mathematics instruction with no supplemental mathematics instruction (the control group). One consideration when using an ITT approach is that the underlying causes for missing data for the intervention group might differ from that for the control group, possibly leading to biased impact estimates. In this study, multiple imputation was used to address missing data and was conducted separately for the intervention and control groups (see the *Missing data* section of this chapter for additional information).

Study timeline

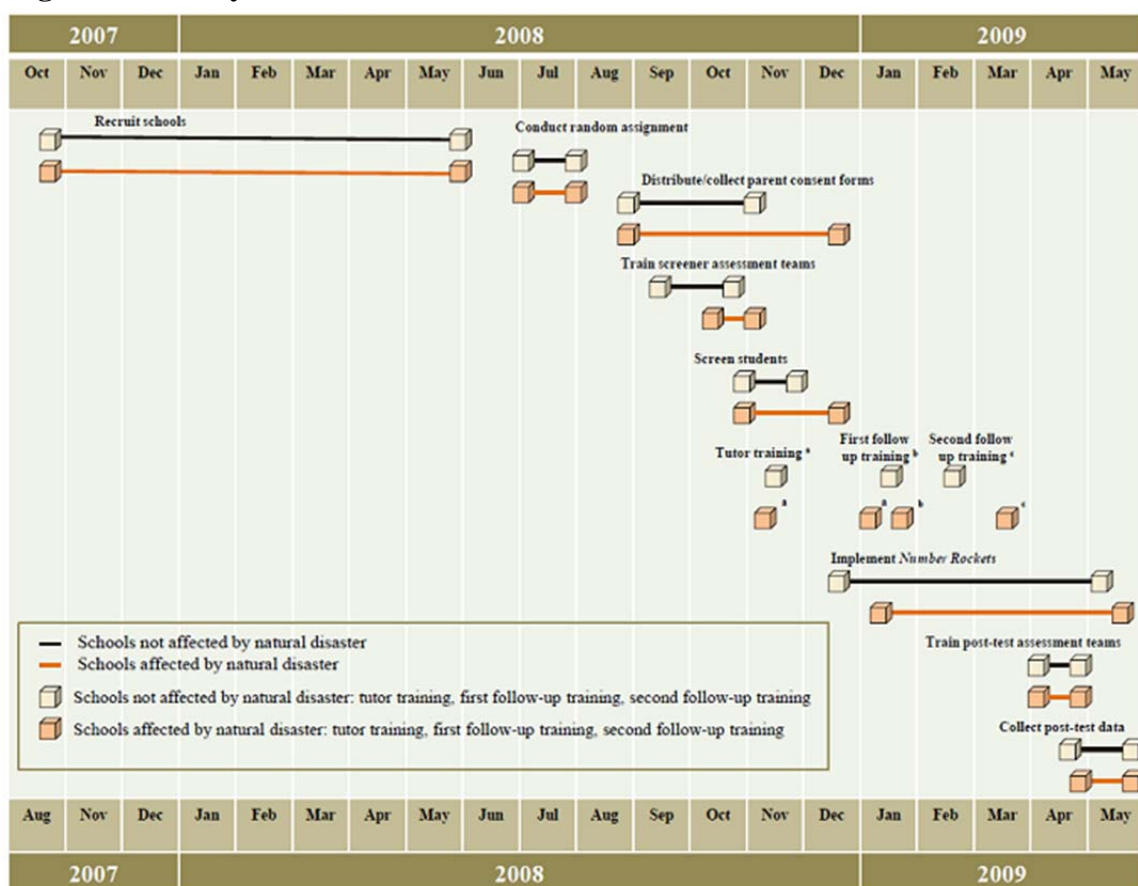
Major study activities spanned from October 2007 through May 2009 (figure 2-1). They included recruitment, random assignment, consent form distribution and collection, screener administrator training, student screening, tutor training (including follow-up trainings), implementation of the intervention, and posttest training and data collection. (See table A-1 in appendix A for a more detailed timeline.)

¹⁸ Due to the lack of other comparable research-validated programs to serve as a counterfactual, and the likely variability across districts in the provision of any local supplemental programs, the decision was made to enroll districts that had no supplemental grade 1 mathematics support in place to control for this potential variability in the counterfactual.

The timelines for 50 of the participating schools were tightly synchronized. As planned, screening for at-risk status in most schools ran from mid to late October 2008 to the first week of November 2008. The implementation of *Number Rockets* began in the first or second week of December 2008 and was completed by the first week of May 2009.

A natural disaster during the consent form collection window affected the timeline for 26 participating schools, causing study schools to close for one or two weeks. After schools reopened, initial consent form return rates were approximately half that of the other schools. The disruption was such that district representatives requested that implementation of *Number Rockets* in affected schools be delayed until the beginning of January 2009. Two more rounds of consent form distributions were then conducted to increase response rates, and at-risk screening continued until the 2008 December holiday break. In addition, to minimize the effects of the two-month lag between tutor training and implementation, the first follow-up training was provided to tutors in these schools two weeks after the start of implementation of *Number Rockets*, as opposed to approximately four weeks after the start of implementation in the other, non-affected schools. Also, to increase the possibility that students in these schools would complete the target minimum 45 lessons, tutors were instructed to negotiate with the principals and teachers for an extra lesson or two each week.

Figure 2-1. Study timeline



- Tutor training for 50 schools was conducted November 13, 17, and 19, 2008. Tutor training for the 26 affected schools was conducted November 10 and 11, 2008, and January 5, 2009.
- The first follow-up trainings for the 50 non-affected schools were conducted January 15 and 20, 2009. The first follow-up training for the affected 26 schools was conducted January 29, 2009—two weeks after the start of implementation of *Number Rockets*—to minimize the effects of the two-month lag between tutor training and implementation.
- The second follow-up trainings for the 50 non-affected schools were conducted February 17 and 19, 2009. The second follow-up training for the affected 26 schools was conducted March 19, 2009.

Source: Study team records collected October 2007–May 2009.

Power analysis and sample size

To determine the sample size, the research team reviewed the outcomes in the research literature for grade 1 mathematics interventions. (For a comprehensive review, see Newman-Gonchar et al. 2009.) The major impetus for the current study was the lack of research on Tier 2 interventions for mathematics in primary grades. The dearth of research, however, inherently meant there was limited research literature on which to base minimum detectable effect size (MDES) estimates.

The statistically significant effect sizes observed in the Fuchs et al. (2005) study (ranging from 0.40 to 0.70) were used as the starting points to conduct power analyses. It was determined that the current study should be powered conservatively to detect an MDES of approximately half the midrange observed in the Fuchs et al. study, and an MDES of 0.30 was targeted to meet this goal. Preliminary power calculations indicated that approximately 60 schools would be needed.

However, because of a better than anticipated response to early recruiting activities, and in part to provide some insurance against school attrition, a minimum target sample size of 70 schools was established¹⁹ and the power analysis was conducted. The power analysis was based on assumptions of the intraclass correlation (ICC) ranging from 0.10 to 0.15, the correlation between the pretest and posttest ranging from 0.30 to 0.70, and the average number of at-risk students in a school being 10 (see appendix B). Under the range of power analysis assumptions evaluated, a target of 70 schools results in a range of MDES values from 0.15 to 0.27.²⁰ As a result, using the most conservative assumptions, an MDES of 0.27 was established for this study. See appendix B for more details on the power analysis.

Target district population and recruitment process

Several criteria determined district and school eligibility to participate in the study. District size and free and reduced-price lunch (FRPL) participation rates were the first criteria considered. Only medium and large districts (14 or more elementary schools with 3 or more grade 1 classrooms per school) with FRPL participation rates of 40 percent or greater were contacted. The purpose of the size criterion was to minimize the total number of districts in the study required to achieve the necessary number of

¹⁹ The target number of schools selected ($n = 70$) also accounted for expected differences in the fidelity of implementation of *Number Rockets* given the design of the current study as an effectiveness (as opposed to efficacy) trial. Constraints on the target sample size included the study budget, a limitation of district recruitment to four or five sites, and other issues involving logistics, personnel, and other resources. Response to recruiting activities continued to be strong; therefore, 76 schools were eventually enlisted and retained in the present study.

²⁰ The What Works Clearinghouse has established an MDES = 0.25 as being of practical significance, even if significant findings are not observed within a study (Institute of Education Sciences 2011).

schools;²¹ the purpose of the FRPL criterion, to increase the likelihood that an adequate number of at-risk grade 1 students would be identified.²² Districts that met these criteria were evaluated for accessibility, such as whether they were within reasonable driving distance of a major airport.

The first step in recruitment was to enlist a national market research firm to identify districts that matched the study criteria; 55 in the REL Southwest Region were identified. To rank order the list, mathematics specialists in the state departments of education in all five REL Southwest states were contacted; these individuals identified districts on the list that, in their estimation, were seeking solutions in early mathematics and would most likely be interested in participating. In one state, in addition to the recommendation of the state mathematics specialists, a regional service center identified promising district contacts for that state. The districts were ranked on the following criteria:

- Number of elementary schools with at least three grade 1 classrooms.
- Number of elementary schools with 40 percent or greater FRPL participation rate.
- Prior working relationship with REL Southwest.
- Known to state officials to be seeking solutions in early mathematics.

The study team intended to contact the districts in order of ranking; however, after just 34 had been contacted, 11 had expressed interest in participating. These 11 had more than 150 schools that met both of the minimum criteria, more than twice the number required based on the power analysis. It was then decided that the 21 remaining districts would not be contacted; it was likely they could not participate because of limits on the number of schools that could enroll.

The 11 interested districts were asked to meet two additional criteria—having a districtwide core mathematics curriculum and no Tier 2 mathematics intervention in place. These two criteria served to reduce variability in Tier 1 mathematics instruction across schools and to ensure the control condition did not include any supplemental mathematics instruction. The study team then met with the interested districts to discuss the study and answer questions. In states where more than one district was interested, the number of elementary schools that met the size and FRPL criteria and the district's commitment (for example, a district's rapid response to communication), were used to choose the district that would participate .

²¹ Because each district has its own research application policy and leadership team to engage, it would be logistically difficult to manage a large number of small district sites.

²² Since 1996, the National Assessment of Educational Progress has used FRPL as a proxy for student poverty (U.S. Department of Education n.d.); students who are economically disadvantaged are considered at-risk for academic difficulties (Arnold and Doctoroff 2003; Roza, Guin, and Davis 2008).

The district sample selected for study participation and inclusion in the random assignment pool consisted of four districts—one each from four of the five REL Southwest Region states.²³ Given a high demand for *Number Rockets* in the Region, including only one district from each state allowed for as much regional participation as possible; however, no attempt was made to collect a representative sample for this study. Table 2-1 summarizes the recruitment activity.

Table 2-1. Recruiting summary data across all districts

Steps in obtaining final district sample	Number of districts
Identified as eligible	55
Total contacted	34
Total declined	10
Total did not reply	13
Total accepted	11
Included in random assignment pool	4

Source: Study team records collected November 2007–May 2008.

In some districts, study participation by all schools that met the size and FRPL criteria was mandated by district administration; in others, participation of eligible schools was voluntary, and all interested schools (46 percent of eligible schools) participated. District-level meetings with school principals and district administrators were held to explain conditions of participation, and district administrators signed formal letters of agreement with an understanding that there was approximately a 50 percent chance that at-risk students would receive tutoring at no cost. No other incentives were provided, except that the *Number Rockets* materials were donated to the participating districts at the end of the study. After the formal agreements for each district were completed, 82 schools were candidates for inclusion in the study sample. Because the target school sample size was 70, this potential school sample provided insurance against attrition.

Matched-pair design and random assignment of schools

Schools were randomly assigned to a condition using a matched-pair design, which increased the probability of baseline equivalence of schools—and the targeted at-risk students within those schools—in both conditions. This option was chosen because there was substantial variability in FRPL participation rates and mathematics achievement between schools in the same district. To create the pairs, schools were

²³ All three eligible districts in the fifth Southwest Region state were contacted, and one responded with possible interest before recruiting efforts were halted. This district is not included in the count of 11 districts expressing interest because REL Southwest had already recruited the district to participate in another large-scale RCT and therefore, excluded them from participation in this study.

matched using FRPL participation rates (using grade 1 student data from the 2007/08 academic year) and mean school mathematics achievement (using scaled scores from the state achievement test for grades 3 and 5 for the previous three school years—2005/06, 2006/07, and 2007/08). Within-district z -scores were calculated for both components; for mean school mathematics achievement, six individual z -scores were created and averaged for the scaled scores of interest. The school composite score was the average of the school's within-district mathematics achievement z -score and its within-district reverse coded-FRPL z -score.²⁴ In each district, schools were sorted on this composite from lowest to highest. The first pair was formed from the two schools in the district with the lowest composite scores, the second was formed from the next two schools, and so on.

Thirty-nine matched pairs of schools (78 of the 82 candidates for inclusion in the study) were created and placed in the random assignment pool as study participants. Four schools could not be paired; these schools were not placed in the random assignment pool nor were they considered study participants. Reasons included an odd number of participating schools within districts and district request for random assignment to be blocked within feeder pattern²⁵ to ensure treatment status would be distributed more evenly across the district.²⁶ Once the 78 schools were paired, the MicrosoftTM Excel RAND() function²⁷ was used to randomly assign one pair member to the intervention condition and the other to the control condition. Potential bias was minimized by having the randomization sequence implemented by REL Southwest staff not directly involved in delivery of the intervention.

After random assignment, but before screening, a control school attrited, causing this pair to be dropped.²⁸ The final analytic sample consisted of 76 schools in 38 pairs (table 2-2).

²⁴ As the composite score was constructed, the proportion of students in each school eligible for FRPL was converted to a within-district z -score. This score was reverse coded so that low proportions of FRPL students were represented by a high z -score. The two z -scores were then averaged for each school.

²⁵ A feeder pattern is a method of organizing the flow of students from elementary and middle schools to specific high schools.

²⁶ Although the four unpaired schools were not placed in the random assignment pool nor considered formal study participants, they were randomly assigned in a separate process to either the intervention or control group. Students in schools assigned to the intervention condition received screening and tutoring identical to study schools, because the study team had committed to providing all interested schools a 50 percent chance of receiving *Number Rockets* (regardless of odd numbers or feeder patterns) to facilitate district enrollment in the study.

²⁷ The RAND() function generates a random number greater than or equal to 0 or less than 1.

²⁸ The intervention member of this pair received services but was otherwise excluded from the study.

Table 2-2. Participating school sample across all districts

<i>Steps in obtaining final analytic sample</i>	<i>Number of schools</i>
Total recruited	82
Paired ^a	78
Randomly assigned ^b	78
Attrited ^b	2
Analytic sample ^b	76

a. Four schools could not be paired (and thus could not be included in the study) for several reasons as described in the text.

b. Because schools were matched, reported numbers are evenly balanced across intervention and control conditions. After random assignment but before consent form distribution, the principal of a control school withdrew that school from the study; the paired-intervention school received services as promised but was no longer considered part of the study.

Source: Study team records collected November 2007–November 2008.

There are costs and benefits of using a matched-pair design. One benefit is that matching schools on FRPL participation rates and school mathematics achievement increases the likelihood that the experimental groups exhibit baseline equivalence after random assignment. This comes at a cost; degrees of freedom are lost in the final impact estimate because of explicit specification of the matched pairs of schools. Exploring this tradeoff, chapter 4 reports the results of a sensitivity analysis that excludes explicit specification of the matched pairs from the model.

Target student population and consent process

The target student population was grade 1 students at risk for difficulties in mathematics who received mathematics instruction in English in a regular education classroom. Because *Number Rockets* requires students to miss non-mathematics regular classroom instruction while being tutored (that is, the intervention is a *pull-out* model where students are removed from the regular classroom), this is considered a potential risk to students. Therefore, active parent consent²⁹ was required for student participation. Principals distributed the consent forms³⁰ to classroom teachers, who distributed them in regular education classrooms where mathematics instruction was in English. Study procedures called for teachers not to distribute consent forms to students in classrooms where a bilingual or non-English mathematics curriculum was used or students in special education classrooms. Because the school districts managed consent form distribution,

²⁹ *Active parent consent* is a requirement that parents return a signed consent form that describes the risks and potential benefits to participation. Parents had to explicitly agree to their child's participation by checking a "Yes" box and signing the form. Active consent differs from typical district implementation and was required by the study's Institutional Review Board.

³⁰ Information provided to parents included details about screening, the intervention, random assignment, and data collection; see appendix C for the English-language version of the consent form. A Spanish version was also prepared, and teachers determined which version of the consent form to send home with each child.

information on the number of students in classrooms who did not meet the inclusion criteria (those not given consent forms) was not collected.³¹

Completed consent forms were returned to the classroom teacher, who then delivered them to the principal. Principals mailed the consent forms to REL Southwest. If less than 70 percent of consent forms were returned by a particular school or district, a second or third set of consent forms was sent to the school and again distributed by the principal through the classroom teacher. (See table 2-3 for the final consent form return rates.)

Consent forms were distributed to all districts beginning in August 2008 and returns were accepted until October 24, 2008. (See appendix A for district-specific dates.)³² All districts requested as a condition of their participation that they be notified in summer (prior to consent form distribution) of the experimental condition to which their schools would be assigned. While districts and schools knew at the time of consent form distribution whether their school had been assigned to the intervention or control condition, parents did not. All parents received the same consent form regardless of whether their child attended an intervention or control school. (See appendix C for a copy of the consent form.) However, parents who inquired using contact information on the consent form were told in which group their child's school had been placed.

³¹ Not all of the participating districts supplied demographic data for the enrolled grade 1 student population; some supplied demographic data for the consenting student population only. As a result, information about the percentage of students in this study who may have been excluded on the basis of English-language proficiency or receipt of special education services in mathematics is limited.

³² For 26 schools, due to a natural disaster, the consent form deadline was extended to the third week of December.

Table 2-3. Parent consent form return rates for students eligible to receive consent forms, by study condition and school district

	<i>Across all districts</i>
Total number of students eligible to receive consent forms (eligible students) ^a	4,844
<i>Intervention</i>	
Eligible students	2,526
Consent form returned	1,992
Consent granted	1,783
Consent granted of eligible students (percent)	70.6
<i>Control</i>	
Eligible students	2,318
Consent form returned	1,473
Consent granted	1,218
Consent granted of eligible students (percent)	52.5
z^b	13.67
p	< .001 ^c

a. *Eligible students* are the students who the district determined met the consent form eligibility criteria: receiving mathematics instruction in English in the regular classroom. Classroom teachers distributed consent forms to eligible students and did not give consent forms to students who received mathematics instruction in a language other than English or in a setting other than the regular education classroom. Because consent form distribution was managed by the districts/teachers, data on how many students were excluded based on these criteria were not collected.

b. z refers to z -tests of the difference between two proportions (the percentage of students with consent granted of the eligible enrollment for the intervention and control groups).

c. A common significance level for a one-tailed test was selected to be 0.05. Using a Bonferroni adjustment for the five comparisons (dividing 0.05 by five) leads to a significance level of 0.01.

Source: Authors' analysis of study team records collected June 2008–December 2008.

In some districts, consent form return rates were significantly higher at intervention schools than at control schools. This finding raises two questions:

- Why was there a differential return rate?
- Does the differential return rate indicate that bias was introduced in student participation after school random assignment?

One possible reason for this differential consent is that assignment status was known to the schools before consent forms were distributed, and control schools were told they would not receive any incentive to participate in the study. So personnel in intervention schools might have worked harder than control schools to obtain parent

consent; however, no data were collected that would allow this to be evaluated. In addition, because only some districts provided complete student rosters and demographic information before consent forms were distributed, it is not possible to compare all enrolled students with those returning consent forms or with those who returned a form and consented.

To investigate whether the differential return rates indicates possible bias in the consent process, several analyses were conducted to evaluate the baseline equivalence of students in the two conditions, including a comparison of screening scores and demographic variables. Results are presented after the description of the screening process.

Screening process

All students whose parents provided consent (1,783 intervention, 1,218 control) were eligible for screening and were screened if available during the screening timeframe (1,643 intervention, 1,076 control). Screened students were identified as at risk (and eligible for the study) based on composite scores from the screener.

Constructing the screener

The screener used in this study included six subtests selected after extensive background review—three from the Fuchs et al. (2005) study and three from research on valid screening measures in mathematics for grade 1 students (Jordan, Kaplan, Locuniak, and Ramineni 2007; Geary 1993; Baker et al. 2006; Clarke et al. 2006). These subtests are described in more detail in the *Measures* section of this chapter. The subtests were assembled into an assessment booklet designed for individual administration. Sample items and additional details for each subtest are in appendix D.

Identifying at-risk students

For each student, a z -score was calculated for each subtest using the mean and standard deviation across all screened students in the 50 participating schools unaffected by the natural disaster.³³ All subtest z -scores were averaged to form a single composite used to rank all students in the sample; in other words, an average score was created where each subtest z -score contributed equally.

³³ If a subtest administration had missing item-level data, that subtest was dropped from the calculation of the screening battery composite, and the composite was calculated using the remaining subtests. Only twelve students of the total 2,719 students screened had a missing subtest score. (See the *Missing data* section of this chapter for more detail.)

Students in the bottom 35 percent³⁴ of the screening sample³⁵ based on this composite score were identified as at risk (615 intervention, 379 control).³⁶ A single study-wide cutscore across schools was selected as the screening approach to control for variability in mathematics proficiency across schools and districts.³⁷ The 35th percentile was selected because it ensured that some students were identified as at risk in each study school³⁸ given the variability in student proficiency across schools, that the total number of students could be served within the resource constraints of the study, and that the cutscore was consistent with other research on mathematics disabilities (for example, Hanich et al. 2001).

See table D-2 in appendix D for descriptive statistics of the screener.

Baseline equivalence

Baseline equivalence was examined for both schools and students. Schools served as the unit of random assignment, and demographic differences between the schools assigned to the intervention and control groups are examined first in this section. Next, to determine whether the random assignment process resulted in similar groups of students at baseline, both the screener performance and the demographics of all screened students, as well as at-risk students in the intervention and control groups, were compared.

³⁴ The Fuchs et al. (2005, p. 496) study, using a “factor score” computed across *Curriculum Based Measurement (CBM)–Computation, Addition Fact Fluency, Subtraction Fact Fluency, and First Grade Concepts/Applications*, applied the 21st percentile as the at-risk cutscore in that study. Initially, the 25th percentile was selected for the current study; however, due to variability in mathematics proficiency levels across schools, and especially across districts, the 25th percentile resulted in several schools having no students identified as at risk. Therefore, the cutscore was reset to the 35th percentile, consistent with other research (for example, Hanich et al. 2001) and ensuring enough at-risk students in each study school to retain statistical power.

³⁵ Due to delays within 26 schools related to the natural disaster, the screener score corresponding to the 35th percentile needed to be determined before data were obtained for those schools, and therefore only data from the other 50 schools was used. This cutscore was then applied in all 76 schools.

³⁶ Nine students were excluded after screening because they could not comprehend the screener instructions or responded entirely in a language other than English.

³⁷ The Fuchs et al. (2005) efficacy study also used a single study-wide cutscore. This practice is consistent with the use of national norms (representing a single consistent cutscore) by districts, such as are available for the widely used *Dynamic Indicators of Basic Early Literacy Skills* (Good and Kaminski 2002). However, in practice, some local education agencies may use other methods of identifying at-risk students, such as the lowest performing percentage or some fixed number of students on each campus. The present study, however, was not designed to evaluate different screening models; therefore, the use of a single rule applied study-wide controlled for variability in mathematics proficiency across districts and schools.

³⁸ The number of at-risk students identified per school ranged from 1 to 49, with an average of 13.1.

School baseline equivalence

Statistically significant differences were observed between treatment and control schools in terms of race/ethnicity (table 2-4). This finding was consistently identified and statistically significant in all four race/ethnicity comparisons in, including: 78 schools, all grade enrollment ($p < .001$); 78 schools, grade 1 enrollment ($p = .034$); 76 schools, all grade enrollment ($p < .001$); and 76 schools, grade 1 enrollment ($p = .006$). Also, for the 76 schools included in the analytic sample, intervention schools had significantly higher grade 1 mean enrollment than control schools ($p = .046$). No other significant differences were found.

Table 2-4. Baseline equivalence of student demographics for all schools randomly assigned, for all grades combined and for grade 1, for all 78 schools initially assigned, and for the 76 remaining after attrition

	<i>Intervention</i> (n = 39)	<i>Control</i> (n = 39)	χ^2	t	p ^a	<i>Intervention</i> (n = 39)	<i>Control</i> (n = 39)	χ^2	t	p ^a
<i>All grades</i>						<i>Grade 1</i>				
<i>78 schools</i>										
Mean number of students (SD)	525.54 (165.47)	462.92 (172.30)		1.64	.106	88.56 (36.06)	73.85 (30.26)		1.95	.055
Percentage of Title I schools	100	92.3		1.77	.081	— ^c	— ^c			
Mean percentage FRPL (SD)	74.5 (15.63)	73.6 (16.00)		0.16	.802	— ^c	— ^c			
Percentage race/ethnicity ^b			172.34		< .001*			8.70		.034*
American Indian/Asian/Other	1.0	1.0				0.9	0.9			
Black	33.4	32.5				30.7	34.0			
Hispanic	56.5	53.3				57.9	54.4			
White	9.1	13.3				10.5	10.7			
	<i>Intervention</i> (n = 38)	<i>Control</i> (n = 38)	χ^2		p ^a	<i>Intervention</i> (n = 38)	<i>Control</i> (n = 38)	χ^2		p ^a
<i>All grades</i>						<i>Grade 1</i>				
<i>76 schools</i>										
Mean number of students (SD)	518.87 (162.29)	443.53 (167.65)		1.68	.097	87.42 (35.83)	72.26 (28.98)		2.03	0.046*
Percentage of Title I schools	100	92.1		1.44	.155	— ^c	— ^c			
Mean percentage FRPL (SD)	74.7 (15.95)	73.6 (16.20)		0.19	.852	— ^c	— ^c			
Percentage race/ethnicity ^b			182.71		< .001*			12.64		0.006*

	<i>Intervention</i> (n = 38)	<i>Control</i> (n = 38)	χ^2	p^a	<i>Intervention</i> (n = 38)	<i>Control</i> (n = 38)	χ^2	p^a
	<i>All grades</i>				<i>Grade 1</i>			
<i>76 schools</i>								
American Indian/Asian n/ /Other	1.0	1.0			0.9	0.9		
Black	34.3	33.8			31.5	35.6		
Hispanic	55.3	51.5			56.7	52.4		
White	9.4	13.8			10.8	11.1		

* Statistically significant at $p < 0.05$.

Note: Percentages may not sum to 100 because of rounding. *SD* is standard deviation; *t* is the *t*-statistic resulting from a two-sample *t*-test; *p* is the probability level associated with the level of the *t*-statistic or χ^2 ; FRPL is free or reduced-price lunch program.

a. alpha = 0.05, two-tailed.

b. These percentages represent the average within-school percentage for schools within the intervention and control groups. Districts reported race/ethnicity in six categories: American Indian, Asian, Black, Hispanic, Other, and White. A multiracial category was not included, as districts did not report these data. Due to small sample sizes, the American Indian, Asian, and Other categories have been collapsed in this table. Unless otherwise noted, Black includes African American, Hispanic includes Latino, Asian includes Native Hawaiian or Other Pacific Islander, and American Indian includes Alaska Native.

c. Grade 1-specific Title I and FRPL participation data were not available.

Source: U.S. Department of Education, National Center for Education Statistics, Common Core of Data, school year 2008/09.

Baseline equivalence of student screener performance

To determine whether random assignment resulted in similar groups of students at mathematics proficiency baseline, the screener performance of all screened students, as well as at-risk students, in the intervention and control groups were compared. There were no statistically significant differences on screener performance between screened students in the intervention and control schools, nor between at-risk intervention and control students in the analytic sample (table 2-5).

Table 2-5. Baseline equivalence of screener scores for all screened students and students identified as at risk, by condition and across all districts

	<i>All screened students</i>				<i>At-risk students (analytic sample)^a</i>			
	<i>Intervention</i> (n = 1,643)	<i>Control</i> (n = 1,076)	<i>t</i>	<i>p</i>	<i>Intervention</i> (n = 615)	<i>Control</i> (n = 379)	<i>t</i>	<i>p</i>
	<i>M</i> (<i>SD</i>) ^b	<i>M</i> (<i>SD</i>) ^b			<i>M</i> (<i>SD</i>) ^b	<i>M</i> (<i>SD</i>) ^b		
Composite screener mean score	−0.04 (0.80)	−0.01 (0.78)	−1.06	0.289	−0.87 (0.38)	−0.85 (0.38)	−0.75	0.458
Subtest								
Quantity Discrimination	22.02 (9.88)	22.61 (9.88)	−1.53	0.127	13.99 (7.25)	14.21 (7.01)	−0.47	0.454
CBM–Computation	6.90 (4.19)	6.88 (4.01)	0.14	0.888	3.59 (3.14)	3.82 (3.21)	−1.13	0.259
First-Grade Concepts/Applications	12.07 (4.41)	12.04 (4.39)	0.18	0.853	8.14 (2.63)	8.15 (2.63)	1.26	0.207
Number Knowledge Test	15.56 (5.36)	15.93 (5.56)	−1.75	0.080	11.45 (3.21)	11.48 (3.21)	−0.54	0.586
Story Problems	4.65 (2.52)	4.78 (2.43)	−1.30	0.194	2.70 (1.99)	2.72 (2.00)	−1.20	0.231
Digits Backward	2.19 (1.57)	2.24 (1.60)	−0.72	0.470	1.02 (1.23)	1.04 (1.23)	0.15	0.880

Note: CBM = Curriculum-Based Measurement; *M* is mean; *SD* is standard deviation; *t* is the *t*-statistic resulting from a two-sample *t*-test; *p* is the probability level associated with the level of the *t*-statistic. All subtest means and standard deviations reported are based on subtest raw scores. The composite mean and standard deviations are based on averaged subtest *z*-scores.

a. The sample of at-risk students evaluated in this study.

b. Subtest standard deviations are not adjusted for clustering.

Source: Authors' analysis of data collected October 2008–December 2008.

Baseline equivalence of students' demographic variables

To determine whether random assignment resulted in similar groups of students at baseline in terms of demographic characteristics, select demographic characteristics of students in the intervention and control groups were also compared by conducting Chi-square (χ^2) tests on district-provided demographic data. A statistically significant difference in the distribution of race/ethnicity between the intervention and control groups for all screened students was observed; however, this was not observed in the at-risk analytic sample, the *subsample* of all screened students on which this evaluation was based (table 2-6). Based on these findings, it is possible that there might be nonequivalence between experimental groups at baseline on unobserved characteristics.

Table 2-6 Baseline demographic characteristics for all screened students and students identified as at risk, by condition and across all districts

Characteristic	All screened students				At-risk students (analytic sample)			
	Condition		χ^2	p	Condition		χ^2	p
	Intervention (n = 1,643)	Control (n = 1,076)			Intervention (n = 615)	Control (n = 379)		
Gender								
Female	49.5	50.9	0.50	0.480	47.0	50.9	1.45	.229
Race/ethnicity ^a			11.14	< .001			0.90	.344
American Indian/Asian/Other	1.1	1.0			1.0	1.1		
Black	39.3	39.3			44.1	43.8		
Hispanic	41.5	47.8			46.7	44.3		
White	18.1	11.9			8.3	10.8		
FRPL								
Yes	30.5	30.0	0.07	0.794	36.0	31.9	1.67	.197
IEP								
Yes	5.6	6.2	0.10	0.721	8.1	7.7	0.07	.787

Note: Percentages may not sum to 100 because of rounding. FRPL is free or reduced-price lunch program; IEP is Individualized Education Program; *p* is the probability level associated with the level of the χ^2 -statistic. Numbers in intervention and control columns are percentages; all χ^2 results are Mantel-Haenszel Chi-Square.

a. Districts reported race/ethnicity in six categories: American Indian, Asian, Black, Hispanic, Other, and White. A multiracial category was not included, as districts did not report these data. Because of small sample sizes, the Asian, American Indian, and Other categories have been collapsed in this table. Unless otherwise noted, Black includes African American, Hispanic includes Latino, Asian includes Native Hawaiian or Other Pacific Islander, and American Indian includes Alaska Native.

Source: Authors' analysis of district-provided demographic data collected May 2008.

Contamination, crossovers, and student mobility

The risk of contamination³⁹ between conditions was low because schools were used as the unit of assignment and classroom teachers were not explicitly provided information on *Number Rockets*. However, several safeguards were used to further minimize this risk. First, *Number Rockets* was implemented at the school level and schools agreed, as a condition of participation, that no other supplemental mathematics instruction outside the regular classroom would be provided in grade 1.⁴⁰ Second, no intervention services were provided to the control schools, nor were study materials or details of the content of *Number Rockets*. Also, in the intervention schools, classroom teachers were not provided any *Number Rockets* materials, ensuring that their whole-class instruction was not influenced. Finally, tutors were instructed not to discuss with teachers either details of tutoring or specifics of student performance.

Also, because schools recruited for this study were required to have at least 40 percent of students eligible for FRPL, a population that typically has nontrivial mobility rates (see for example Xu, Hannaway, and D’Souza 2009), a plan to deal with mobility patterns was developed before screening to account for possible student crossover.⁴¹ At-risk students who moved between study schools were posttested as members of the group to which they were originally assigned. Students ($n = 5$) who moved from one intervention school to another continued to receive *Number Rockets*. They were added to a tutoring group on a lesson closest to the last lesson they had received in the previous school. Students participating in *Number Rockets* who moved from an intervention school to a control school in the district (fewer than 5) were not provided tutoring services after the move. However, because this study used an ITT approach, these students were still considered part of the intervention group. Conversely, a student moving from a control school to an intervention school ($n = 0$) would have been considered a control student. There was little possibility that a classroom teacher could have acquired the *Number Rockets* materials and conveyed them to the student. However, if a student moved within the district to a non-study school or left the district entirely ($n = 75$), no attempt was made to posttest the student. Missing TEMA–3 scores were estimated as described in the *Missing data* section of this chapter. (See appendix E for more information about student mobility.)

³⁹ In this study, *contamination* refers to potential instances where teachers or students assigned to the control condition receive intervention details or strategies.

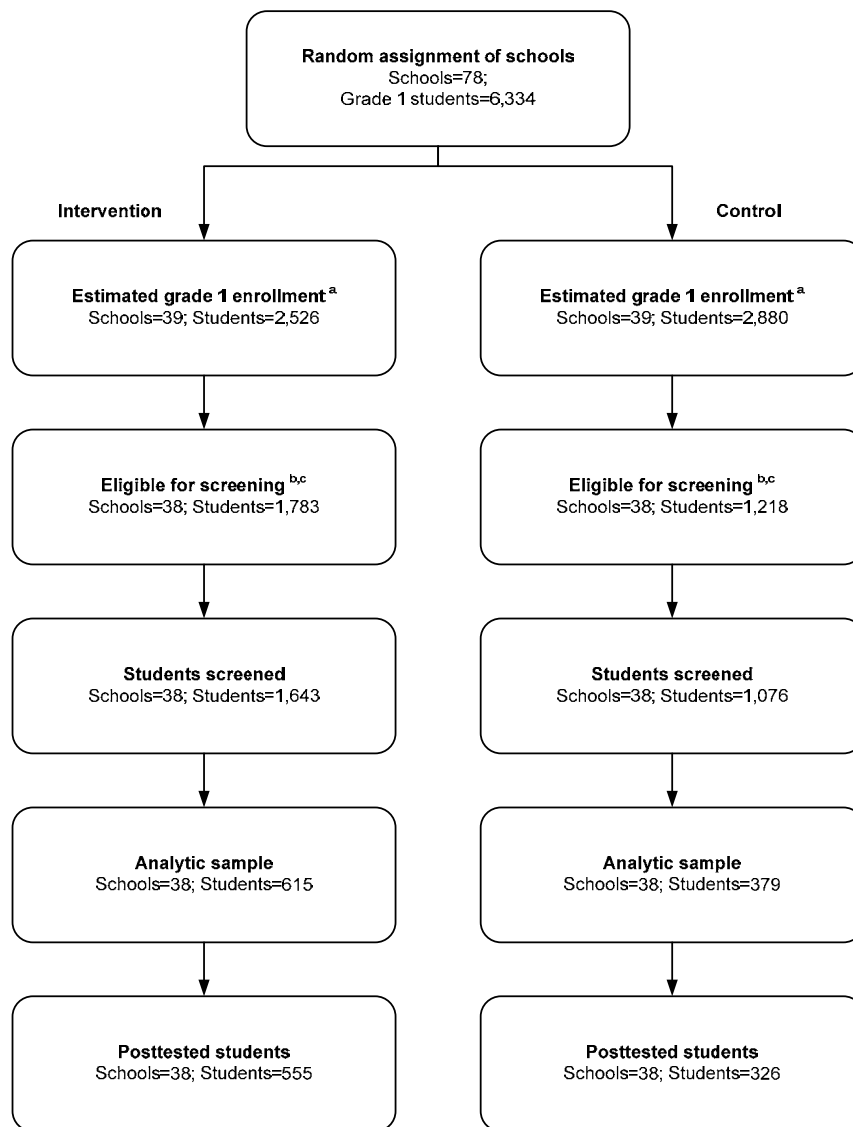
⁴⁰ Due to resource constraints, study personnel could not verify if all schools adhered to this agreement.

⁴¹ In this study, *crossover* refers to intervention students moving to control schools or control students moving to intervention schools at any point during the implementation.

Sample at each phase of the study

The sample retained in each phase of the study for schools and students is detailed in figure 2-2. This figure includes sample sizes for several groups of students:

- *Estimated grade 1 enrollment.* Not all districts provided complete student rosters; these estimated total enrollment counts were derived from the Common Core of Data (U.S. Department of Education n.d.) and included students receiving special education services and students not receiving mathematics instruction in English.
- *Eligible for screening.* Teachers distributed consent forms to the target population of grade 1 students; that is, students who received mathematics instruction in English in a regular education classroom. Grade 1 students who received consent forms and later received parent consent to participate were “eligible for screening.”
- *Students screened.* Data were collected for grade 1 students who were administered the study screener. Some students ($n = 282$), absent during all attempts to assess during each district’s two-week screening window, were not screened.
- *Analytic sample.* Grade 1 students scoring in the bottom 35 percent of the screened students were identified as at risk and participated in the study.
- *Posttested students:* Posttest data were collected for grade 1 students who participated in the study and who were enrolled in a study school at the time of posttest data collection.

Figure 2-2. Schools and grade 1 students in the sample for each phase of the study

Note: Not all districts provided complete student rosters. Therefore, the number of grade 1 students at random assignment and the estimated grade 1 enrollment figures were approximated using data from the National Center for Education Statistics Common Core of Data (n.d.).

- a. Classroom teachers distributed consent forms only to the target population of grade 1 students (those who received mathematics instruction in English in a regular education classroom), a subset of the estimated enrollment.
- b. After random assignment but before screening, a control school withdrew, causing the pair to be dropped. Students in the control school were not screened; students in the intervention school were screened solely to determine at-risk status; this was done to provide the appropriate students with *Number Rockets* regardless of inclusion in analytic sample.
- c. Students who received mathematics instruction in English in a regular education classroom and received parent consent were eligible for screening.

Source: National Center for Education Statistics n.d.; study team records collected June 2008–May 2009.

Measures

Several types of measures were used in this study. Described in this section are the six subtests that comprise the screener, the measures used to assess fidelity, and the measures used to evaluate study outcomes for the confirmatory and exploratory research questions.

Screener measures

The screener used in this study consisted of six individually administered subtests. All six subtests measure a construct that prior research has indicated is predictive of success in first grade mathematics outcomes. All six subtests have been used in prior studies where students were screened for at-risk status. Most of the six are similar to measures commonly used as fall benchmark assessments or progress monitoring tools in response-to-intervention models. Three subtests were used in the Fuchs et al. (2005) study and the other subtests were selected from research on valid screening measures in mathematics for grade 1 students. Greater detail for each subtest follows.

The three measures selected from the Fuchs et al. (2005) study⁴² are:

- *Curriculum-Based Measurement–Computation* (Fuchs, Hamlett, and Fuchs 1990). This measure is a one-page test displaying 25 items that sample the grade 1 computation curriculum. Students have 2 minutes to complete as many problems as possible. The score is the number of problems answered correctly and, because it is not a norm-referenced measure, is provided as a raw total score. The Fuchs et al. (2005) study reported a coefficient alpha⁴³ of 0.95.
- *First-Grade Concepts/Applications* (Fuchs, Hamlett, and Fuchs 1990). This grade 1 measure is a three-page test with 25 items sampling typical grade 1 concepts/applications, including numeration, concepts, geometry, measurement, applied computation, charts/graphs, and word problems. With this measure, the assessor reads the words in each item aloud. For 22 items, students have 15 seconds to respond before the assessor reads the next item; for the remaining 3, students have 30 seconds per item. The score is the number of correct answers and, because it is not a norm-references measure, is provided as a raw total score. The Fuchs et al. (2005) study reported a coefficient alpha of 0.92 for that sample.

⁴² The three measures: *Curriculum-Based Measurement–Computation*, *First-Grade Concepts/Applications*, *Story Problems* were found to be sensitive to the *Number Rockets* program in that study. Effect sizes for these three individual measures were reported as 0.40, 0.67, and 0.70, respectively, when comparing change scores (posttest minus pretest) of intervention and control groups. In the current study, however, these measures were used as screeners only and were not re-administered at post-test.

⁴³ Coefficient alpha is a measure of the internal consistency or reliability of an assessment.

- *Story Problems* (originally Jordan and Hanich 2000). The version of this measure used in The Fuchs et al. (2005) study consisted of 14 items and is most appropriate for grade 2. A more recent eight-item version that is more appropriate for grade 1 students (Jordan et al. 2007) was substituted for use in the present study. The assessor read each item aloud. The student was instructed to respond verbally to several types of story problems including *combine*, *change*, *equalize*, and *compare*. Each item followed a similar format. For example, “John has 6 pennies and Lauren has 3 pennies. How many pennies do they have altogether?” Each of the problems included simple addition and subtraction of 9 or less (Jordan et al. 2007). This test is not norm-referenced and provides only a raw total score. Jordan et al. (2007) did not report a reliability coefficient (although the Fuchs et al. [2005] study reported a coefficient alpha of 0.86 for the 14-item version). Jordan, Kaplan, Oláh, and Locuniak (2006) used the same version of the test that Jordan et al. (2007) used, at four time-points across the school year, and reported coefficient alphas from 0.58 to 0.77. Jordan et al. (2007) did report six validity coefficients collected across the school year between this measure and a composite formed from the Calculation and Applied Problems subtests of the Woodcock-Johnson—Third Edition Letter/Word subtest (WJ—III Letter/Word; Woodcock, McGrew, and Mather 2001) collected at the end of the school year. These correlations ranged from 0.47 to 0.62, increasing on average as the school year progressed.

These three measures were combined with three others: a broad measure of number sense, a key construct that has been identified in more recent research as predicting success in early mathematics (Baker, Gersten, and Lee 2002; Jordan et al. 2007); a measure of magnitude comparison, generally considered a key component in beginning number sense (for example, Booth and Seigler 2008; Gersten, Jordan, and Flojo 2005); and a measure of working memory, where poor performance has been observed for students identified with mathematics disabilities (for example, Wechsler 2003). The three additional measures are:

- *The Number Knowledge Test* (Baker et al. 2006). This subtest has been used to chart children’s developmental profiles of numerical competency (Case et al. 1996) and to study the effect of mathematics instruction on kindergarteners from low socioeconomic status families (Griffin 1997). As conceptualized by the developers, the Number Knowledge Test provides a general measure of number sense. The test contains four levels of increasing difficulty. Students advance to the next level if they complete a certain number of items on their current level. The final score is derived by summing the scores from each level completed. Item response theory reliability for the Number Knowledge Test with a sample of 470 students (Baker et al. 2006) was estimated at 0.93 using a one-parameter model and at 0.94 using a two-parameter model. The Number Knowledge Test has a

predictive validity coefficient of 0.72 (Baker et al. 2006) from spring of kindergarten to the end of grade 1 with the Stanford Assessment Test Series—Ninth Edition mathematics subtest (Harcourt Assessment, Inc. 2004).

- *Quantity Discrimination* (Clarke et al. 2006). This measure assesses a student's ability to make magnitude comparisons. The ability to make numerical judgments of magnitude is a key cornerstone of beginning number sense. In this individually administered measure, students examine pairs of numbers between 0 and 10 and identify which is greater. There are a total of 56 items (pairs of numbers) and the assessment ends after five consecutive incorrect or no responses or after 60 seconds. Concurrent and predictive validity coefficients range from 0.64 to 0.67 on the Stanford Early School Assessment Test—Second Edition (Harcourt Brace Educational Measurement 1996), an early-grade form of the Stanford Achievement Test Series, Tenth Edition.
- *Digit-Span Backward* (Geary 1993). This is a measure of auditory working memory that first appeared in Binet and Simon's (1905/16) original work on intelligence. An individually administered measure, it requires students to listen to a string of digits presented by the examiner (2–8 digits in length) and verbally repeat the string in the reverse order. There is evidence (for example, Wechsler 2003) that difficulties performing this task are associated with the development of mathematics disabilities.

Additional subtest details and descriptive statistics for the screener subtests are in appendix D.

Fidelity of implementation measures

Three types of fidelity measures were collected during *Number Rockets* implementation. *Lesson fidelity checklists* were coded from audio recordings of tutoring sessions. *Instructional logs* were used to track administrative information about each tutoring group session. In addition to these two fidelity of implementation measures, *Classroom instruction checklists* were used to measure the fidelity of schools' adherence to the developer's instruction to use *Number Rockets* as a strictly supplemental mathematics program.

Lesson fidelity checklists

The Fuchs et al. (2005) study developed a series of lesson-specific fidelity checklists to evaluate tutors' administration of the scripted *Number Rockets* lessons. These checklists consisted of 8–30 specific tutor verbal actions checked against audio recordings of the tutoring sessions. Because the checklists had not been developed for all lessons, REL Southwest developed additional lesson fidelity checklists as needed, using

the same design principles; these were used in addition to the original checklists to assess fidelity of implementation.

In this study, tutors were instructed to capture audio recordings of each tutoring session.⁴⁴ For each tutor, four *Number Rockets* sessions were evaluated (if audio recordings were available) using lesson fidelity checklists. Tutors were unaware of which lessons would actually be scored and coded. The first evaluated session was used to provide background information for the coaches⁴⁵ who worked with the tutors, not to examine tutor fidelity (discussed in chapter 3). The other three were used to judge fidelity of implementation.⁴⁶ Additional information about the lesson fidelity checklists in in chapter 3, and one lesson fidelity checklist used is in appendix F.

Instructional logs

After each lesson, tutors were required to report the following information to the study team: tutoring group number, lesson completed, date of lesson, session time in minutes, any comments about the lesson and students, and absenteeism. All instructional log data were collected at the individual lesson level and were aggregated by study staff and reported at the group level to determine the average number of lessons delivered. See table F-1 in appendix F for a sample of the aggregated instructional log data.

Classroom instruction checklists

Per district agreement, teachers were to ensure that tutors did not remove students in the intervention group from regular mathematics instruction. To explore whether schools adhered to this guidance, a classroom instruction checklist was developed. While gathering the students for the day's tutoring session, tutors completed the checklist by asking teachers to describe all instructional activities that students would miss during the session; the classroom instruction checklist was completed each day a lesson was delivered during the specified one-week window. One week was selected as the unit of time to collect these data because some activities, such as physical education, music, and art vary on a daily basis. This one-week sample does not necessarily represent the instruction missed by students throughout the intervention period, but it provides descriptive information on the instructional tradeoffs for students receiving this intervention in a real-world effectiveness implementation.

⁴⁴ Tutors were trained in the use of audio recorders during their initial training and submitted their recordings every one or two weeks.

⁴⁵ Refer to chapter 3 for more information about the role of coaches.

⁴⁶ Ten percent of the audio sessions were evaluated by two independent reviewers; using checklist total scores as the unit of analysis, the interrater reliability, as measured by an intraclass correlation, was 0.65 across tutors and sessions.

A single classroom instruction checklist (a sheet) captures missed instruction for one group for one day. Space for recording up to three students for that group was provided to account for the possibility that students were drawn from more than one classroom. The classroom instruction checklist was collected during the last month of tutoring. Either three or four sheets were collected for each group during a single week, depending on the number of tutoring sessions scheduled. See figure F-2 in appendix F for a copy of the classroom instruction checklist.

Outcome measures

The TEMA–3 was used to answer the confirmatory research question and exploratory research questions 1 and 3. The WJ–III Letter/Word subtest was used to answer exploratory research question 2.

Test of Early Mathematics Ability–Third Edition

The TEMA–3, an individually administered mathematics test, was used as the primary outcome measure. The TEMA–3 assesses a broader set of mathematics skills than those represented in the pretest screener measures. Given that state mathematics assessments were not available as grade 1 outcome measures, the TEMA–3 was selected because, as an individually administered test, it was appropriate for the grade level of the students and measured mathematics achievement broadly, as state accountability measures do. The TEMA–3 was not considered as a pretest since it is unlike the universal screening measures that would typically be used by districts, and its length and individually administered format also would have made it cost prohibitive to do so.

The TEMA-3 test measures both formal and informal mathematics skills, with items sampling the following topics: numbering, number comparisons, calculation, concepts, numeral literacy, number facts, and calculation. The content assessed by the TEMA–3 is designed to be consistent with typical grade level curricula taught in schools (Ginsburg and Baroody 2003). The reliability of the measure is reported ($\alpha = 0.95$; test-retest = 0.82–0.93), and norms are based on a sample weighted to be nationally representative and scaled to a mean of 100 and a standard deviation of 15. Test administration takes about 30 minutes.

An individually administered test was selected for this study because they are more likely to elicit optimal performance from students who are inexperienced test-takers (Sattler and Hoge 2006). The TEMA–3 is designed to be sensitive to student abilities in the lowest quarter of the distribution of grade 1 mathematics skills (Ginsburg and Baroody 2003).

Assessment teams were recruited locally and trained by study staff experienced in administering individual assessments and in training others to administer them. (See appendix A for training dates.) Assessors had experience administering individual

assessments or working directly with grade 1 students. Teams of three to eight assessors worked in each school. All eligible students in a school were typically assessed in one or two days. Students sick or absent on scheduled assessment dates were typically assessed within one week of the originally scheduled date.

Woodcock-Johnson–Third Edition Letter/Word subtest

To examine the possible influence on reading proficiency of reducing the classroom time devoted to reading instruction, data were collected at posttest using the WJ–III Letter/Word subtest (Woodcock, McGrew, and Mather 2001). A commonly administered measure of letter and word identification, it takes less than five minutes for administration. Students are shown lists of letters, then words of increasing difficulty on a page-by-page basis. Students read the letter or word aloud and are given one point for each letter or word read correctly until the student responds incorrectly or not at all to six consecutive numbered items) or until the end of the test. The reliability of the WJ–III Letter/Word subtest is $\alpha = 0.98$ for six-year-old students (Woodcock, McGrew, and Mather 2001). Speece et al. 2004 reported a correlation of 0.81 between the Test of Early Reading Achievement–Second Edition (Reid, Hresko, and Hammill 1989), a general test of early reading ability, and the Woodcock-Johnson Psychoeducational Battery–Revised Letter/Word subtest (Woodcock and Johnson 1990), demonstrating that the previous version of the WJ–III Letter/Word subtest correlates with a measure of general reading ability. The WJ–III Letter/Word subtest was administered during posttest assessment immediately after the TEMA–3, and data from this administration is in the discussion of exploratory analysis 2 in chapter 5.

Data analyses

This section summarizes the data analysis methods. It overviews the hierarchical linear modeling (HLM)⁴⁷ approach used to evaluate the confirmatory research question and the corresponding sensitivity analyses. It also addresses the methods used to evaluate the exploratory research questions, including the sensitivity analysis conducted for exploratory research question 1. (All models are provided in appendix G.) The section concludes by discussing the method used to deal with missing data.

⁴⁷ Hierarchical linear modeling (or multilevel modeling) is a form of linear regression analysis which accounts for clustering effects for hierarchical data by simultaneously modeling variance within and between levels. The variance between clusters stems from the fact that student scores from the same school will tend to be more similar to each other (as compared with scores from other schools) because students are sharing the same instructional context with their school or cluster.

Confirmatory analysis

The confirmatory research question was evaluated using HLM models⁴⁸ that compared TEMA–3 outcomes of students in the intervention schools with TEMA–3 outcomes of students in the control schools. HLM was used to account for the nested structure of the data (meaning students were nested in schools that were nested in pairs of schools). Specifically, a three-level HLM model was constructed with students at level 1, schools at level 2, and school pairs at level 3. Other models could have been used; however, because of the large number of school pairs and the fact that the research question did not focus on estimating specific pair-level effects, a parsimonious model was selected with school pairs modeled as a separate level with a random effects model. Note, however, that there was no intention to generalize findings beyond the current sample of schools. Chapter 4 describes the results from a sensitivity analysis with a two-level HLM model that excludes specification of the school pairs entirely. Each student's screener composite score was included as a pretest covariate in the model to obtain higher statistical precision of the parameter estimates (Bloom, Richburg-Hayes, and Black 2007; Raudenbush, Martinez, and Spybrook 2005). There were no observed statistically significant differences between the analytic sample of students in the intervention and control groups at baseline (see tables 2-5 and 2-6). Only one outcome measure (the TEMA–3) and a single confirmatory impact analysis were proposed; therefore, correction for multiple comparisons was not necessary.

In addition to the statistical significance of the *Number Rockets* effect, the analysis gauges the magnitude of the impact with the effect size index. For HLM, the appropriate method for calculating effect size is Hedges' *g* (Institute of Education Sciences 2008); for more detail, see appendix G.

Sensitivity analyses

Six sensitivity analyses were conducted to provide additional context for the confirmatory analysis.

The first analysis evaluated whether the impact estimate was sensitive to the exclusion of 26 schools from the sample. Because of complications stemming from the natural disaster, the schedule for the affected schools differed from that of the other 50 schools.⁴⁹ The HLM model for the confirmatory analysis was applied to a sample including only the 50 non-affected schools.

The second sensitivity analysis evaluated the robustness of the confirmatory impact estimate to the decision to explicitly specify the matched school pairs in the HLM

⁴⁸ The study team used *HLM 6.02 for Windows* (Raudenbush, Bryk, and Congdon, R., 2006)

⁴⁹ The 26 affected schools began *Number Rockets* implementation in January 2009, with four or five lessons delivered per week wherever possible. The 50 non-affected schools began *Number Rockets* implementation in December 2008, with three lessons typically delivered per week.

model. Using matched pairs involves a possible tradeoff, where statistical power may be gained by evaluating impacts between pairs of similar treatment and control schools, at the cost of losing degrees of freedom in the HLM analysis (which can reduce statistical power). A two-level HLM analysis was conducted, adapting the HLM model used for the confirmatory analysis without the level specifying school pairs (level 3).

The third sensitivity analysis evaluated the robustness of the model to the inclusion of the pretest covariate; the analysis was conducted by modifying the HLM model used for the confirmatory analysis to exclude the pretest covariate.

The fourth sensitivity analysis evaluated the sensitivity of the study results to the chosen missing data approach (multiple imputation); the sensitivity analysis was conducted using casewise deletion, an alternate missing data approach in which only students with complete TEMA-3 scores were included in the analysis. The confirmatory HLM model was used but with a smaller sample size ($n = 881$).

The fifth sensitivity analysis examined the robustness of the confirmatory impact estimate to the decision to include students not identified as at risk ($n = 45$) in some tutoring groups, in order to satisfy the study's commitment to provide *Number Rockets* to a minimum of nine students at each intervention school; the analysis was conducted using the confirmatory HLM model, excluding at-risk students ($n = 970$ retained, $n = 24$ excluded) who were assigned to tutoring groups that included students who were not part of the at-risk analytic sample. In other words, entire student groups that included students not at risk, were excluded.

The sixth sensitivity analysis also examined this decision by using the confirmatory HLM model and excluding entire school pairs in which any tutoring groups included students who were not part of the at-risk analytic sample. In this analysis, 29 of 38 pairs were retained, resulting in $n = 883$ students retained for the analysis and $n = 111$ students excluded.

Exploratory analyses

The first exploratory question focused on the relationship of baseline mathematics proficiency with the impact of *Number Rockets*. The main impact model, modified to include a cross-level interaction between school treatment status and student pretest screening scores, was used to examine whether the treatment impact differs as a function of baseline mathematics proficiency. A sensitivity analysis was also conducted, using the screener composite scores to split students into three ability groups (low, medium, high), to evaluate the robustness of the exploratory finding to this alternate approach. The impact estimate for each ability-grouped student subsample was estimated separately using the same HLM model as the confirmatory analysis. In the subsequent sensitivity analysis, these three impact estimates were tested against each other to determine if ability-grouping resulted in statistically different impact estimates.

The second exploratory question focused on the possible influence of reduced classroom reading instruction time on reading proficiency. The same analytic model for the main impact was used here, too; however, WJ–III Letter/Word scores were used as the outcome variable.

The third exploratory question focused on the possible variation of the intervention effect of *Number Rockets* on student TEMA–3 performance, based on the average number of tutoring sessions delivered to tutoring groups at each intervention school. The same analytic model was used as for the confirmatory impact estimate, with the addition of the implementation variable (number of sessions) and the interaction between the implementation variable and the treatment indicator at the school-pair level. Results for all three exploratory questions are reported in chapter 5 and the models are described in appendix G.

Missing data

This study used an ITT approach to analyze data. The ITT approach makes no assumptions about student participation after the offer to participate; it is an estimate based on those *offered* services. Students may choose not to participate in the intervention, may move outside the research site, or may not receive the intervention with the intended fidelity.

An ITT analysis provides an impact estimate for the entire sample of students and thus must address any missing outcome data in that sample. Multiple imputation⁵⁰ was used to obtain an ITT impact estimate. In particular, multivariate stochastic sequential regression-based multiple imputation was used whenever students were missing posttest scores. Multiple imputation was chosen for the following reasons:

- Multiple imputation provides asymptotically unbiased estimates when the missing data mechanism is missing at random.
- Multiple imputation does not reduce the statistical power like other methods, such as case-wise deletion.
- Multiple imputation provides appropriate standard errors to account for the imputation process, which would not be available given a single imputation.

See chapter 4 for the results of the fourth sensitivity analysis, conducted to examine whether the study findings are sensitive to the missing data approach.

⁵⁰ Multiple imputation (Little and Rubin, 1987) is an approach where missing data values are estimated using existing data values. This is done independently more than one time and the results combined to provide appropriate estimates of the standards errors for these missing values (or estimates of the uncertainty in estimating missing values).

Multiple imputation was conducted separately for the intervention and control groups.⁵¹ Five multiply imputed datasets were created for each group and combined to create five overall imputed datasets.⁵² The imputation model for each multiply imputed dataset included the screener composite score, gender, race/ethnicity, FRPL status, IEP status, English language learner status, and dummy indicator variables for schools and school pairs to account for the clustered structure of the data.

All students were required to have a non-missing screener composite score. Therefore, students who could not be screened after repeated attempts in a two-week window⁵³ were ineligible for the study. If a subtest administration resulted in missing item-level data, it was dropped from the calculation of the screening battery composite, and the composite was calculated using the remaining subtests.⁵⁴ Multiple imputation was used to estimate TEMA–3 posttest scores for 113 students (60 intervention, 53 control), including those who could not be assessed because they relocated to a nonparticipating school within the district or relocated out of the district ($n = 75$), were absent due to extended illness or were otherwise not available during the posttest window ($n = 28$), or had an invalid assessment, were removed from *Number Rockets* participation by a parent or legal guardian, or from regular mathematics instruction to attend special education programs ($n = 10$). For the students removed from regular mathematics instruction to attend special education, data collected or associated with the implementation of *Number Rockets* were not used or provided to any individual to assess, identify, recommend, or suggest evaluation or referral to special education services or subsequent removal from the study. Posttest data were not collected for these students and whether the students were responsive to the mathematics intervention is unknown; no interim assessment before posttest was collected.

While it cannot be known if there was differential attrition based on unobserved variables, the two samples of students with complete TEMA–3 scores and those with missing TEMA–3 scores were compared on observed demographic variables and the screener composite score (table 2-7). Based on this analysis, there were no statistically significant differences between students with and without missing posttest scores. The results show no evidence that processes causing missing TEMA–3 scores were nonrandom but do not provide information about unobserved variables. As discussed, multiple imputation provides unbiased estimates when the missing data mechanism is missing completely at random. Chapter 4 provides additional information on this issue,

⁵¹ Only students that were part of the analytic sample were included in the multiple imputation. Therefore, students at the two schools that attrited prior to screening were not included.

⁵² Other researchers such as Graham et al (2007) have recently proposed that combining even larger numbers of imputed data sets result in more statistical power.

⁵³ In some cases, typically due to long-term illness, a third week was allowed to screen any remaining students.

⁵⁴ Of the 2,719 students screened, 12 had one missing subtest score. All missing scores could be traced back to examiner error

where the results of a sensitivity analysis conducted for the confirmatory impact analysis are presented, using only students with non-missing data.

Table 2-7. Demographic characteristics and mean screener composite scores for students with TEMA–3 scores and for students missing TEMA–3 scores

	<i>TEMA–3 score status</i>		χ^2	p
	<i>Complete</i>	<i>Missing</i>		
Sex				
Female	49.0	42.3	0.92	.338
Race/ethnicity ^a			3.63	.057
American Indian/Asian/Other	1.0	-- ^c		
Black	44.4	40.7		
Hispanic	46.1	43.4		
White	8.5	-- ^c		
FRPL				
Yes	34.9	31.0	0.67	.415
English language learner				
Yes	12.0	9.7	0.51	.476
IEP status				
Yes	8.1	7.1	0.13	.717
<i>Screener composite</i>				
	<i>Mean (SD)^d</i>	<i>Mean (SD)^d</i>		
	–0.86 (0.38)	–0.91 (0.40)	–1.45	.149

FRPL is free or reduced-price lunch program; IEP is Individualized Education Program; *p* is the probability level associated with the level of the χ^2 statistic.

Note: Demographic characteristics of the students for whom TEMA–3 scores were available are reported in percentages; all χ^2 results are Mantel-Haenszel Chi-Square.

a. Districts reported race/ethnicity in six categories: American Indian, Asian, Black, Hispanic, Other, and White. A multiracial category was not included, as districts did not report these data. Due to small sample sizes, the American Indian, Asian, and Other categories have been collapsed in this table. Unless otherwise noted, Black includes African American, Hispanic includes Latino, Asian includes Native Hawaiian or Other Pacific Islander, and American Indian includes Alaska Native. Percentages may not sum to 100 because of rounding.

b. TEMA–3 scores were missing for 11.4 percent of students in the analytic sample (113 of 994), including 9.8 percent of intervention students (60 of 615) and 14.0 percent of control students (53 of 379). Missing TEMA–3 scores were imputed as described in text. A two-tailed *z*-test of the difference in attrition proportions for each experimental group was conducted with $\alpha = 0.05$ and was not statistically significant ($z = 1.933$; $p = .053$).

c. These two cells were suppressed because one of the cells represented fewer than 3 cases and is, therefore, a disclosure risk.

d. Screener composite standard deviations reported here are not adjusted for clustering.

Source: Authors' analysis of study team data collected April 2009–May 2009.

Chapter 3: Implementing *Number Rockets*

This chapter describes how *Number Rockets* was implemented and presents results on the fidelity of implementation. The *Number Rockets* intervention, including a sample lesson, is described first. A discussion of the *Number Rockets* tutor training and a description of the tutors follows. The chapter concludes with an analysis of the fidelity of implementation data collected and information on the cost of the intervention.

Description of *Number Rockets*

Number Rockets is a supplemental intervention implemented by a tutor with a group of two or three students, typically meeting three times per week for approximately 17 weeks. Students meet with the tutor outside the classroom during regular school hours, usually around a small table, for about 40 minutes per tutoring session. The tutor follows instructions and reads text aloud from a lesson script that includes highly prescribed feedback and prompting procedures to use with students as they perform various individual and group activities. For the last 10 minutes of the session, the tutor works with the students on mathematics fact practice using flashcards. The tutor prepares a deck of flashcards for each student prior to each lesson based on the student's current skill level with addition and subtraction facts.⁵⁵ The flashcards' difficulty increases with the skill of individual students and are independent of the group progression through the lessons. Throughout each lesson, tutors also use a behavior management system, representing an established protocol of tutor behaviors intended to maintain student attention on, or redirect student attention to, *Number Rockets* tasks. Consistent with the scripted nature of *Number Rockets*, the behavior management system minimizes variability in student discipline and positive reinforcement practices that individual tutors might otherwise use in absence of a standardized protocol.

In the behavior management system, students receive award points for reaching mastery criteria for a lesson and at various timed intervals averaging five minutes each⁵⁶ when all members of the group are "on task" (defined as listening carefully, working hard, and following directions). When a student accumulates a predetermined number of points and completes a points sheet maintained by the tutor, the student is allowed to choose a small reward (such as a small toy car, keychain, or pencil eraser). Most students earned a reward approximately every two sessions.

The *Number Rockets* program covers 17 topics, each divided into three to six lessons, not all of which are required. There are 63 potential lessons included in the program. The entire program, however, can be completed in as few as 41 lessons. One

⁵⁵ Examples of addition and subtraction facts included on the flashcards are 1+1, 2+1, 3+1, or 5-0, 4-0, 3-0.

⁵⁶ The points sheet includes specific predetermined time intervals that ranged from two to nine minutes. These intervals average five minutes.

complete lesson is delivered each day that student groups meet. If the entire group of students meets the mastery criteria for a topic during a required lesson, the additional lessons for the topic are skipped. Students still cover all 17 topics regardless of the number of lessons skipped due to meeting mastery criteria. (The topic and lesson sequences are detailed in appendix H.)

Sample Number Rockets lesson

To illustrate the general structure of the lessons and the types of activities in which students engage, a portion of a lesson from Topic 6, *Introduction to Place Value*, is presented in figures 3-1 and 3-2. (For the Topic 6 sample lesson in its entirety, see appendix I.) Topic 6 consists of three required lessons delivered over three separate days. On day 1 students review content taught as part of the previous topic (in this example, Topic 5). Next, the tutor begins the lesson and provides an explicit example and feedback as students work through the example set of numbers. This process is repeated for 14 more numbers and the session concludes with mathematics fact practice.

Topic 6, Day 1 lesson excerpt

After reviewing material covered in the previous topic, the tutor begins Topic 6, Day 1, by presenting the concept of *place value* verbally and by writing an example (figure 3-1). Students are each provided a worksheet on which to write answers for the rest of the lesson. Both positive feedback (such as “great work” and “that’s right”) and corrective feedback (such as “these numbers are different from each other because . . .”) to students is scripted (Paulsen and Fuchs 2005, p. 57). After the tutor introduces a verbal example of place value, the tutor demonstrates one way to represent place value with fingers (figure 3-2).

The lesson continues with the introduction of Base-10 blocks and repeated practice of translating 14 more numbers using the blocks. If the next lesson for this topic was not required, students would complete an additional worksheet to check for content mastery. If the mastery criterion was met by all members of the group, any additional lesson(s) for this topic would be skipped.

The final 10 minutes of each tutoring session consists of mathematics fact practice with each student’s flashcards. The tutor works with one student at a time, while the other students watch. The tutor gives the flashcards to the selected student for one minute, and the student responds to as many flashcards as he or she can in that period, taking as much time as needed for each card. If a student responds to one of the cards incorrectly, the tutor leads him or her through a hand-counting procedure⁵⁷ to answer the problem. Once

⁵⁷ Part of the *Number Rockets* flashcard activity involves a counting procedure using both hands. For a problem such as $5 + 3$, students are instructed to make a fist with one hand to represent the larger number (in this example, the number 5), and to represent the smaller number with the other hand by extending the corresponding number of fingers. (In this example, the number 3 would be represented by extension of the thumb, index, and middle fingers.) Students are then instructed to cross arms at the wrist, with the

the minute is up for that student, the tutor continues around the circle, taking turns with each student. A second round is conducted, where each student attempts to answer more cards correctly than in the first round. If the lesson takes longer than planned, the flashcard activity is truncated to keep the total session time within approximately 40 minutes.

Figure 3-1. Excerpt from Topic 6, Day 1 lesson *Number Rockets* script: tutor introduces place value

Great work. Today we're going to be working on place value.

Write the numbers 5 and 13.

How are these numbers different?

If the student gives an incorrect response say, These numbers are different from each other because the 13 takes up two places, but the 5 only takes up one place. How are the numbers different?

Students should respond something like:

5 takes up one place

13 takes up two places

That's right. These numbers are different from each other because the 13 takes up two places; two numbers together make up 13. But 5 only takes up one place. So, 5 takes up one place, but 13 takes up two places.

Give students Topic 6 Day 1 Tutoring Sheet 1.

These places have a special name. Write 5 in the ones place of the first box on Topic 6 Day 1 Tutoring Sheet 1. This (point to the 5) is called the ones place. Five only has one place. Write 5 on your sheets. Show me 5.

Source: Paulsen and Fuchs 2005, p. 57.

extended finger hand over the fist hand. Students are instructed to bump their wrists together while verbally counting from the larger number (5) and folding in one extended finger at a time while counting (in this case, saying 6, 7, 8) until all fingers are folded and the correct answer is reached. In this example, the student would bump his or her wrists together four times (once to represent the 5, then each time one of the three extended fingers is folded).

Figure 3-2. Excerpt from Topic 6, Day 1 lesson *Number Rockets* script: tutor models place value

Write 13 in the second box. Look at 13. In 13, the 3 is in the ones place. (point to the 3 in 13). Every number has something in the ones place. But, look, 13 takes up two places. (point to the 1 in 13). This is called the tens place.

Now I'm going to show you how to show 13 with your fingers. When we have a number that's in both the ones and tens place we'll "flash" all 10 fingers and then count the ones. Let me show you what I mean.

Flash 10 and count up 11, 12, 13.

Now you show me 13 with your fingers.

Great work.

Source: Paulsen and Fuchs 2005, p. 58.

Tutor training

In this section, the development of the tutor training and support program for *Number Rockets* is described first, followed by the tutor training activities and tutor performance review process. A discussion of the district coaches is next. The demographic characteristics of the tutor population, details of tutor attrition, and a description of tutoring assignments conclude the section.

Development of Number Rockets tutor training

Because the present study implemented *Number Rockets* in four large urban districts, a formal training and support program was required for the tutors.⁵⁸ So the study team worked with the *Number Rockets* developers (members of the Fuchs et al. [2005] study team) to create a support program similar to the professional development and training support provided by publishers of other curriculum products. For example, in a recent large-scale Institute of Education Sciences (IES) evaluation comparing Tier I curricula in early mathematics (Agodini et al. 2009, pp. 26–28), curriculum publishers were allowed to specify (and provide) the level of support for each program. For the four programs described in Agodini et al. (2009), one or two days of initial training were

⁵⁸ Because the Fuchs et al. (2005) study was an efficacy trial implemented by the *Number Rockets* developers, formal training and professional development materials for the tutors had not been developed. Instead, that research team met with the tutors weekly and conducted personal observations of performance. In addition, individual coaching feedback was provided on an ad hoc basis.

provided, with at least one follow-up meeting between teachers and trainers during the school year.⁵⁹

The initial collaboration with the *Number Rockets* developers was a two-day meeting. The developers provided members of the current study team with an overview of the critical elements of *Number Rockets* that needed to be presented during the initial training of new tutors. Preliminary training materials were then developed by the current study team and reviewed by the *Number Rockets* developers. Feedback from the developers focused on three elements:

- Defining the critical components of the program, such as how students meet mastery criteria.
- Ordering the core concepts of the program to be introduced and practiced.
- Integrating opportunities for tutors to practice during the training day.

Tutor training in practice

Tutors in the current study were provided with a one-day introduction to *Number Rockets*, with two two-hour follow-up sessions approximately 4 and 10 weeks later. Training was provided by district coaches (described in the *Role of the district coach* section that follows) who responded to questions about the program by phone or email throughout the study, as needed. The initial one-day tutor training session included four elements:

- Overview of the scope and sequence of *Number Rockets*.
- Review of all necessary materials needed to implement the intervention.
- Discussion of the intervention's critical elements.
- Practice time for some of the more difficult aspects of the lessons with feedback from the coaches and their fellow tutors.

(Additional details of tutor training are provided in appendix J.)

Tutor performance review

After implementation began, digital audio recordings of intervention sessions were reviewed for each tutor to evaluate their performance. As described in chapter 2, tutors were instructed to capture audio recordings of each tutoring session and were

⁵⁹ In Agodini et al. (2009), publishers for three of the programs specified a one-day initial training; for the other program, the publisher specified a two-day initial training. Subsequently, teachers in two of the programs reported 2.2 to 2.9 days of follow-up training, and teachers in the other two programs reported 0.4 to 0.5 days of follow-up training. All four programs provided phone and email support for the entire school year.

provided a small digital audio recorder. The Topic 2, Day 1 lesson was selected for the initial review of each tutor's performance.⁶⁰ Two members of the study team individually reviewed the audio file and assigned a fidelity score, based on the lesson fidelity checklist.⁶¹ Each tutor was given an overall rating of *pass*, *pass with support*, or *fail*. Tutors received a rating of pass if 70 percent or more of the desired behaviors were observed. For all tutors scoring less than 70 percent, the next available audiotaped lesson was evaluated by the study team as a second fidelity check. If the second check was below 70 percent, the tutor was assigned a rating of fail and released from the study; if the second fidelity check was above 70 percent, the tutor was assigned a rating of pass with support. District coaches supported the tutors assigned a rating of pass with support by reviewing the lesson by phone, pointing out program components that were not delivered effectively and providing strategies for student behavior management when necessary.

Role of the district coach

The study team provided follow-up support for effective implementation of *Number Rockets* using district coaches. Three members of the study team were designated as such and each was assigned to one of the four districts. (One district coach was assigned to two districts.) Each district coach had extensive experience teaching elementary school, in addition to other school- and district-level experience working with students who struggled in school. The district coaches averaged 30 years of experience (from 13–46 years) working as educators (teachers, coaches, and/or administrators).

The district coaches became experts with *Number Rockets* by participating in a two-day training session with the intervention developers, followed by practice with grade 1 student volunteers on several key lessons. Each district coach then conducted the initial one-day training for the assigned district(s). In each case, the initial training was observed by one or both of the other district coaches, who provided constructive feedback. Throughout the study, the district coaches had access to two *Number Rockets* developers by email and phone.

The district coaches served as the main contact with tutors working in the district. The coaches, responsible for answering questions arising during the day-to-day implementation of *Number Rockets*, supervised activities in their assigned district(s). In addition, approximately 4 and 10 weeks after the intervention began, the coaches

⁶⁰ Because some lessons were skipped if all students in the group met certain mastery criteria on required lessons, the Topic 2, Day 1 lesson used to screen tutors was either the second, third, or fourth lesson delivered by a new tutor. See appendix H for a list of lessons organized by topic and day.

⁶¹ Subsequently, the two study team members discussed their results until consensus was achieved for each item on the checklist; no score differences remained after consensus meetings. For more information on these lesson fidelity checklists, see chapter 2, the *Fidelity of implementation* section later in this chapter, and appendix F.

conducted the two follow-up training sessions at each site.⁶² The trainings addressed common questions and issues arising during the intervention and provided a refresher on its critical elements. A district coach led the follow-up meetings by verbally reviewing common questions and concerns. Typically, these were procedural issues such as how to present flashcards, changes in techniques, formats the students would use to solve certain types of mathematics problems, use of accompanying materials, behavior management, and pacing of the lessons to ensure that lessons were completed within the allotted time.

Description of tutors

Tutors were recruited and hired through a nationwide temporary services company that handled all human resources functions for the study team.⁶³ Candidates were required to be certified teachers and/or have experience in the elementary (K–5) grades. Typical applicants were retired teachers or teachers active in the substitute teacher pool in the local district. Preference was given to those with experience teaching in individual or small-group settings.

The percentage of tutors who were female was 85.3. In terms of racial/ethnic distribution, 46.7 percent of the tutors were White, 42.7 percent were Black, and 8.0 percent were Hispanic. The percentage of tutors who reported a Master’s degree or higher was 32.0 percent, and 44.0 percent reported having 0–5 years of teaching experience. The percentage of tutors who reported they were retired teachers and/or worked as a substitute teacher was 77.3. See table K-1 in appendix K for more detailed demographic information for the final tutor sample.

Tutor attrition

Prior to intervention implementation, 84 tutors were trained on *Number Rockets* administration, including extra tutors hired in anticipation of tutor attrition during implementation (table 3-1). In addition, 27 new tutors were hired to replace tutors who were released or withdrew during implementation.⁶⁴ If a tutor was released or withdrew, the students assigned that tutor were reassigned to another tutor and/or another tutoring

⁶² Tutors who missed a follow-up training, or were hired after a follow-up training, participated in either a two-hour face-to-face meeting or a conference call with the district coach. During the call, district coaches covered the same content, questions, and issues that arose during the face-to-face sessions. These follow-up training calls lasted between 30 and 120 minutes, depending on the issues raised and the number of participants.

⁶³ All tutoring activities were carried out by study staff; no school personnel were paid as part of this study. Schools provided minor administrative support but did not receive any monetary compensation.

⁶⁴ New tutors were provided the one-day training delivered by the assigned district coach and the two two-hour follow-up training sessions. In some cases, new tutors hired later in the study participated in their first follow-up training during the same group session that was other tutors’ second.

group. An effort was made to maintain lesson progression for each student and group, regardless of tutor assignment changes.⁶⁵

Table 3-1. Tutor training activities and attrition

	<i>Total number of tutors</i>
<i>Initial training: 8 hours</i>	
Participated	111
Released after initial training ^a	4
Withdrew after initial training	6
<i>First follow-up training: 2 hours^b</i>	
Number of active tutors	101
Participated	98
Released after first follow-up	7
Withdrew after first follow-up	5
<i>Second follow-up training: 2 hours^c</i>	
Number of active tutors	89
Participated	82 ^d
Released after second follow-up	0
Withdrew after second follow-up	3
<i>Final number of tutors</i>	86 ^e

a. Fidelity implementation checklists were completed within days of the first student session. Tutors who did not meet quality standards on such checklists were released.

b. Across all districts, the first follow-up training was face-to-face, except for four tutors who received a two-hour follow-up training via conference call with the district coach.

c. Across all districts, the second follow-up training was face-to-face, except for ten tutors who participated in a 50-minute conference call with the district coach and seven who had a 30-minute one-on-one telephone call with the district coach. Five tutors hired to replace tutors who withdrew or were released from the study did not receive a second follow-up training because they joined the study late in the intervention period.

d. This number represents the total number of tutors active at any point in the study.

e. The final number of tutors represents the number of active and alternate tutors during the final month of *Number Rockets* implementation. Four of the 86 tutors served as alternate tutors and did not actively participate in implementation of the intervention. An alternate tutor was someone hired and trained to substitute for an active tutor in the event the active tutor was not able to complete intervention implementation. Of the original 86 tutors at the onset of this study, 82 (86-4) were actively tutoring in the last month of the study.

Source: Study team training records from November 2008–May 2009.

⁶⁵ For more detail on the timing of the intervention delivery and all trainings, please refer to chapter 2, figure 2-1 and appendix A, table A-1.

Tutoring assignments

Tutor assignments were designed to minimize the number of different tutors in each school so that individual tutors could become better known to school staff and thus work more effectively. The average number of tutors in each school was 4.3, with a range of 2 to 14 per school. The average number of tutoring groups per tutor was 2.8, with a range of 1 to 5 groups per tutor. Study staff created tutoring groups by assigning two or three students to each group.

Groups consisted of either two students (41 groups; 18.1 percent) or three students (186 groups; 81.9 percent), with an average of approximately 2.8 students per group. Note that multiplying the number of groups by group size results in $n = 640$, while the total intervention sample was 615 students. This is because while all intervention schools were promised that nine or more students would receive *Number Rockets* (a minimum of three tutoring groups), not all schools had nine students identified as at risk by the screener. In other words, some students not considered at risk (but close to the screener cutscore) participated in *Number Rockets* tutoring groups so that a minimum of nine students received services in each intervention school. These extra students were not considered part of the study or the analytic sample, and no attempt was made to posttest them. Two sensitivity analyses were conducted to examine the robustness of the confirmatory analysis to the inclusion of students assigned to tutoring groups with students not part of the at-risk analytic sample; results are in chapter 4.

Tutoring group assignments were based on tutor availability, school schedules, and classroom schedules. For ease of accessibility, students from the same class were assigned to the same tutoring groups whenever possible. If a student moved from the school, leaving a group with only one student, the groups were reorganized so that each group would have two or three students. When this was necessary, students joined the tutoring group most closely aligned in the lesson sequence with their previous group. Tutors retained the same tutoring groups throughout the study when possible. Due to tutor turnover and the need to reconstitute groups when students moved from a school, approximately 34% of student groups experienced a change in tutor during the study, with approximately 9% of students experiencing more than one change in tutor.

Fidelity of implementation

This section describes three fidelity measures—two used to evaluate tutor fidelity of implementation and one to measure the school fidelity to the developers' guidance on using *Number Rockets* strictly as a supplemental mathematics program. The data collected using these measures provide important contextual information for interpreting the impact estimates (see chapter 4).

The first measure, the *lesson fidelity checklist*, examined the extent to which the intervention was implemented as intended by the developers. The second, the *instructional log*, tracked administrative information for each tutoring group session, such as how many lessons were completed. The third, the *classroom instruction checklist*, is a measure of program implementation by schools, who were to ensure that core mathematics instruction was not missed because of participation in *Number Rockets*. The *classroom instruction checklist* was used to characterize the instructional tradeoff that results from students missing regular classroom instruction to attend mathematics tutoring and, specifically, to determine if students missed regular mathematics instruction to attend *Number Rockets* tutoring sessions. (These measures are in appendix F.)

Lesson fidelity checklists

Lesson fidelity checklists were created by the *Number Rockets* developers to evaluate the implementation of the intervention in the Fuchs et al. (2005) study. The *Number Rockets* developers defined fidelity as the completion of key steps in each lesson of the intervention, and their lesson fidelity checklists consisted of explicit actions that tutors needed to perform, as defined by the lesson script. For example, item number 4 in the lesson fidelity checklist for Topic 15, Day 1 states, “The tutor explains that s/he will use Base-10 blocks to help add $85 + 12$ and that he/she first will show 85 with the blocks” (see figure F-1 in appendix F for the complete lesson fidelity checklist for Topic 15, Day 1). The number of steps on the lesson fidelity checklists ranged from 8 to 30 (mean = 12).

In the current study, tutors were responsible for audio recording each session. Samples of the recordings were evaluated using the lesson fidelity checklists. (Information on lesson fidelity checklist coding is in the next section of this chapter.) When a tutor did not provide an audio recording⁶⁶ corresponding to one of the targeted lesson fidelity checklists, the study team, with guidance from the developers, created a new lesson fidelity checklist for a lesson that was recorded. The new lesson fidelity checklists used the same format and structure as the original checklists and included the critical actions that the tutors needed to complete. Because the number of items on each lesson fidelity checklist varied by lesson, fidelity was calculated as the percentage of steps completed by the tutor.

Training of lesson fidelity checklist coders

Coders were hired to review audio recordings of tutoring sessions and calculate lesson fidelity using the checklists. Each coder had a Master’s degree in education and their teaching experience ranged from 21 to 33 years in elementary or middle school.

⁶⁶ Approximately 25 percent of lessons were not successfully recorded for a number of reasons including: battery failure, memory limitations of the digital audio recorders, ambient noise, and user errors.

Coders were introduced to *Number Rockets* through a brief overview of the program elements, a discussion of a typical lesson that captured the essential aspects of the program, and a segment on the procedure for the flashcard activity. Coders were informed that tutors were asked to read from a script and implement a behavior management system so that instruction was consistent for all students.

After the introduction, coders received training that included practice with one videotaped lesson and three audiotaped lessons to ensure complete understanding of how tutor behaviors corresponded to fidelity checklist items. Coders first viewed a videotaped lesson and practiced coding tutor behaviors. The videotaped lesson was identical to the one used in the initial tutor training and ensured consistency in tutor and coder training. The lesson fidelity checklist identified the key behaviors a tutor should implement when teaching the lesson, and coders were asked to determine whether the behavior was observed. At frequent intervals, the current study team would stop the video to ask the coders if they thought there was evidence of the specific behaviors. The trainers also highlighted specific tutor behaviors and key features of the lesson that met the criteria for adequate implementation.

Once coders had an explicit visual representation of what high fidelity to a lesson meant, they coded three more lessons from audio recordings and the accompanying flashcard activities. After each practice audio lesson, the trainer conducted a debriefing about the training and coders practiced calculating fidelity percentages. Coders had to achieve 85 percent accuracy at the item-level on the practice lessons before coding lessons that were used to calculate tutor fidelity. Before conducting actual audio fidelity checklist evaluations, coders achieved accuracy criteria on four lessons (one video, and three audio). Ten percent of the audio sessions were evaluated by two independent coders; using checklist total scores as the unit of analysis⁶⁷, the interrater reliability was calculated across tutors and sessions (intraclass correlation = 0.65; see appendix L for additional details on coder training).

Lesson fidelity checklist findings

For each tutor, a digital audio recording from an early, middle, and late topic (Topics 1–5, 6–10, and 11–17) in *Number Rockets* was collected and evaluated. Across all districts and lessons, average lesson fidelity was 85.0 percent (table 3.2); at the district level, average lesson fidelity ranged from 81.4 percent to 90.3 percent (district level data not shown in table). Also, across the entire sample, average lesson fidelity increased from *early* (83.3 percent), *middle* (84.1 percent) to *late* (87.6 percent) lessons.

⁶⁷ Note that the use of total scores, rather than item-level scores as the unit of analysis, limits the interpretability of the ICC. Two raters may have reached agreement on total score, while being in disagreement on an unknown number of individual items.

Table 3-2. Fidelity of lesson implementation by district and implementation-period

<i>Across all districts</i>		
<i>Lessons</i>	<i>n</i>	<i>Mean</i>
Early (Topics 1–5)	76	83.3 (18.2)
Middle (Topics 6–10)	88	84.1 (16.4)
Late (Topics 11–17)	82	87.6 (17.2)
All lessons	246	85.0 (17.2)

Note: Means are percentages; numbers in parentheses are standard deviations. The team attempted to sample three lessons for each tutor, one from each third of the intervention period: *early*, *middle*, and *late*. When not possible, a lesson from another point in the intervention period was substituted. It was not always possible to sample three lessons from each individual due to poor audio quality issues. This occurred for seven tutors (8.1 percent).

Source: Audio recordings collected November 2008–May 2009; lesson fidelity checklists coded July 2009–November 2009.

The *Number Rockets* developers did not provide a standard for acceptable fidelity of implementation. However, the Fuchs et al. (2005) study reported an average fidelity rate of 94.6 percent.⁶⁸ Because the current study was an effectiveness trial with less supervision of tutors, a lower level of fidelity was expected.

Instructional logs

Instructional logs (see appendix F for an example) tracked tutoring group activities. Tutoring groups were scheduled to receive a minimum of three lessons for each full week of regular instruction. Tutors were directed to make up lessons missed due to scheduling conflicts or whole-school events (such as school not in session due to inclement weather) and to add additional sessions where possible. In 26 schools, implementing a fourth or fifth session per week was made a high priority to compensate for the delay resulting from the natural disaster. As discussed, 45 lessons was the target for the minimum number of lessons to be delivered.⁶⁹

Across study districts, an average of 48.4 lessons were completed, ranging from an average of 47.0 to 50.8 (table 3-3). The percentage of groups completing 45 or more lessons ranged from 57 percent to 98 percent, with 75 percent of groups across the study completing at least 45 lessons. Data documenting the cause of variability in the number of lessons delivered per tutoring group were not collected systematically; however, some tutors did record related information in their instructional log, indicating that reasons

⁶⁸ The Fuchs et al. (2005) study reported average fidelity of implementation for Topic 4, Day 1 or 2 (early in the intervention) and Topic 16, Day 1 (late in the intervention); fidelity averaged 93.5 percent for Topic 4 and 95.6 percent for Topic 16.

⁶⁹ In the Fuchs et al. (2005) study, all tutoring groups completed 48 lessons.

lessons were cancelled include assemblies, fire drills, inclement weather, and tutor illness.

Table 3-3. Average number of lessons delivered per tutoring group

<i>Number of tutoring groups</i>	<i>Number of lessons delivered</i>					<i>Groups completing 45 or more lessons (percent)</i>	
	<i>Mean</i>	<i>Standard deviation</i>	<i>Minimum</i>	<i>Median</i>	<i>Maximum</i>		
Across all districts	227 ^a	48.4	(6.45)	30	49	66	74.9

a. Groups consisted of either two (41 groups, 18.1 percent) or three (186 groups, 81.9 percent) students, with an average of approximately 2.8 students per group. Note that multiplying the number of groups by group size results in $n = 640$, while the total intervention sample consisted of 615 students. This is because while all intervention schools were promised that nine or more students would receive *Number Rockets* (a minimum of three groups), not all schools had nine students identified as at risk by the screener. In other words, some students not considered at risk (but close to the screener cutscore) participated in *Number Rockets* groups throughout the study so that a minimum of nine students received services in each treatment school. These extra students were not considered part of the study or the analytic sample, and no attempt was made to posttest them.

Source: Authors' analysis of instructional logs completed by tutors November 2008–May 2009.

Across all districts, 32.4 percent of groups completed all 17 *Number Rockets* topics (table 3.4). All groups completed the program through Topic 11 (Addition facts), and more than half (56.8 percent) of the groups completed the program through Topic 16 (Two-digit subtraction).⁷⁰

Although districts varied in the number of lessons completed and the amount of content covered, most tutoring groups completed a majority of the *Number Rockets* intervention (74.9 percent completed at least 45 lessons and 56.8 percent completed at least 16 of the possible 17 topics).

⁷⁰ See appendix H for a complete list of the 17 topics.

Table 3-4. Percentage of tutoring groups that completed intervention Topics 11 through 17

<i>Intervention topic</i>		<i>All groups across all districts (n = 227)</i>
11	Addition facts	100.0
12	Subtraction facts	98.7
13	Addition and subtraction facts review	92.8
14	Place value review	88.7
15	Two-digit addition	81.1
16	Two-digit subtraction	56.8
17	Missing addends	32.4

Source: Author's analysis of instructional logs completed by tutors November 2008–May 2009.

Classroom instruction checklist

The *Number Rockets* intervention is implemented during the school day and students are removed from regular classroom instruction to participate. Therefore, students participating in *Number Rockets* trade additional mathematics instruction for other classroom activities (for example, reading instruction). The study team sought to characterize this tradeoff and collected information on the type of classroom instruction each student missed when attending *Number Rockets* sessions.

The classroom instruction checklist (provided in appendix F) was used during one week in the month before the intervention was completed, to record classroom activities students missed. When tutors picked up their students from the classroom for a lesson, they asked the teacher what instruction or activity the students would miss during the next 40 minutes.⁷¹ This was done for each *Number Rockets* lesson taught that week. Classroom schedules are typically the same each week, and because student sessions were always scheduled at the same time during the week, it is likely that the classroom activities recorded are reasonably accurate.

The classroom activities students missed when in *Number Rockets* sessions are summarized in table 3-5. Tutors were allowed to record multiple activities missed during a lesson. For example, students could miss music and recess. Therefore, the columns in table 3-5 sum to more than 100 percent, and the categories are not necessarily mutually exclusive.

⁷¹ Note that only the instructional activities missed were captured on the classroom instruction checklist; instructional time per activity was not recorded.

Table 3-5. Percentage of reported *Number Rockets* lessons in which a specified classroom activity was missed

	<i>Across all districts (738 lessons reported)</i>
<i>Reading</i>	
Whole class reading instruction ^a	8.8
Guided reading	4.3
Independent work	8.1
Small group reading instruction ^b	4.7
Learning centers	2.2
Other reading	5.7
<i>Language arts</i>	
Spelling	2.2
Writing ^c	10.8
<i>Mathematics</i>	
Whole class mathematics instruction	11.4
Small group mathematics instruction	3.8
<i>Other</i>	
Art	6.6
Computer lab	4.7
Music	6.5
Physical education	9.6
Recess	3.8
Science	10.2
Social studies	8.3

Note: *Lessons reported* refers to each time a tutor removed a group of students from class to deliver a *Number Rockets* lesson during the week that these data were collected. Multiple activities could be recorded as missed during the removal time. Therefore, the columns sum to more than 100 percent, and the categories are not necessarily mutually exclusive.

a. Whole class reading instruction includes phonics, phonemic awareness, fluency, vocabulary, and comprehension.

b. Small group reading instruction includes phonics, phonemic awareness, fluency, vocabulary, and comprehension.

c. Writing includes writing process, grammar, and punctuation.

Source: Authors' analysis of classroom instruction checklists collected May 2009.

The formal agreement established with the districts stated that students would not be removed from core mathematics instruction to participate in *Number Rockets*; the intervention is designed to reinforce regular classroom instruction, not replace it. Still, it was reported that 11.4 percent of the classroom activities missed during the week that was sampled were whole-class mathematics instruction. In fact, the most common single activity missed overall was whole-class mathematics instruction, followed by instruction in writing and science.

Cost of implementation

A cost estimate for implementing *Number Rockets* was not available from the developer. The program was developed in an academic research setting and is not currently available through a commercial publisher. But the tutor lesson scripts and student worksheets are available for public sale from the copyright holder (Vanderbilt University). For this study, the research team adapted these master documents into a tutoring kit bag including multiple tutor binders, laminated cards, flashcard sets, Base-10 blocks, clipboards, timers, student folders (including points sheets), student rewards, and a complete copy of all student worksheets for each intervention student.

Study staff estimated the annual average cost of providing *Number Rockets* as implemented in this study at approximately \$10,000 per school and \$700 per student, including the cost of materials, coaches, tutors, and ongoing support. These cost estimates are based on a single initial year of intervention implementation. Costs might be somewhat lower in subsequent years as tutors become experienced with the program and need less coaching support. Also, many *Number Rockets* materials can be reused. These estimates include no costs associated with screening, as it is considered a Tier 1 activity.

Chapter 4: Estimating the impact of *Number Rockets* on student achievement

This chapter examines the effectiveness of *Number Rockets* as a Tier 2 mathematics intervention by presenting empirical evidence on whether grade 1 students at risk for mathematics difficulties in the intervention group scored higher on the Test of Early Mathematics Ability—Third Edition (TEMA–3; Ginsburg and Baroody 2003) than grade 1 students at risk for mathematics difficulties in the control group. First, to empirically demonstrate that the baseline equivalence of the intervention and control groups (accomplished through random assignment) was maintained at posttesting, the two groups in the analytic sample were compared on both demographics and baseline screener scores. Second, to address the confirmatory research question on the effect of *Number Rockets* on grade 1 mathematics achievement, findings estimating this effect are presented. The results of sensitivity analyses are also presented.

Maintenance of baseline equivalence

No statistically significant differences were found at baseline between the analytic sample of students in the intervention group and students in the control group on the observed demographic characteristics (see table 2-6). In addition, no statistically significant differences were found at baseline for the groups on mean screener composite scores (see table 2-5). Due to student mobility and student absence, TEMA–3 scores were not available for 11.4 percent (113 of 994) of students in the analytic sample: 9.8 percent (60 of 615) of intervention students and 14.0 percent (53 of 379) of control students. The overall and differential attrition rates observed in this study meet the What Works Clearinghouse’s category of *Meets Evidence Standards*, defined as “attrition is expected to result in an acceptable level of bias even under conservative assumptions” (Institute of Education Sciences 2011).⁷² Analyzing students for whom TEMA–3 data were collected showed no statistically significant differences between intervention and control students on the observed demographic characteristics or mean screener composite scores (table 4-1).

⁷² The What Works Clearinghouse’s *Meets Evidence Standard* for attrition is based on two criteria: differential attrition rate (of < ≈8 percent at maximum) and the overall attrition rate (of < ≈45 percent at maximum) (Institute of Education Sciences 2011). Acceptability is based on the relative rates of these two criteria.

Table 4-1. Demographic characteristics and mean screener composite score for students with TEMA–3 scores, by assigned condition

	<i>Condition</i>		χ^2	p
	<i>Intervention</i>	<i>Control</i>		
<i>Sex</i>				
Female	47.0	52.5	2.42	0.120
<i>Race/ethnicity^a</i>			1.29	0.257
American Indian/Asian/Other	0.9	1.2		
Black	45.2	42.9		
Hispanic	46.1	46.0		
White	7.8	9.8		
<i>FRPL</i>				
Yes	36.6	31.9	1.98	0.160
<i>IEP</i>				
Yes	8.3	7.7	0.11	0.744
<i>Screener composite</i>	<i>Mean (SD)^c</i>	<i>Mean (SD)^c</i>	t	p
	−0.86 (0.38)	−0.84 (0.38)	0.90	0.367

Note: Percentages may not sum to 100 because of rounding. Demographic characteristics of the students for whom TEMA–3 scores were available are reported in percentages; all χ^2 results are Mantel-Haenszel Chi-Square; IEP is Individualized Education Program.

- a. Districts reported race/ethnicity in six categories: American Indian, Asian, Black, Hispanic, Other, and White. A multiracial category was not included; districts did not report these data. Due to small sample sizes, the American Indian, Asian, and Other categories have been collapsed in this table. Unless otherwise noted, Black includes African American, Hispanic includes Latino, Asian includes Native Hawaiian or Other Pacific Islander, and American Indian includes Alaska Native.
- b. TEMA–3 scores were not available for 113 of 994 students (11.4 percent) in the analytic sample, which were imputed as described in the text. Missing TEMA–3 scores were observed for 9.8 percent (60 of 615) of intervention students and 14.0 percent (53 of 379) of control students.
- c. Standard deviations of the screener composite reported here have not been adjusted for clustering of students within schools.

Source: Screener data collected October 2008–November 2008; TEMA–3 data collected April 2009–May 2009.

Confirmatory research question findings

Number Rockets caused a statistically significant 4.28 point difference on TEMA–3 scores favoring the intervention group over the control group ($p < .001$). This 4.28 point difference corresponds to an effect size of 0.34 standard deviations on the TEMA–3 (table 4-2).

Table 4-2. Impact of *Number Rockets* on mathematics achievement of grade 1 students as measured by the TEMA–3, by assigned condition

<i>Outcome measure</i>	<i>Intervention</i> (<i>n</i> = 615)		<i>Control</i> (<i>n</i> = 379)		<i>Estimated intent-to-treat impact</i>			
	<i>Mean</i>	<i>Standard deviation</i>	<i>Mean</i>	<i>Standard deviation</i>	<i>Estimated impact</i>	<i>Standard error</i>	<i>p</i>	<i>Effect size^a</i>
TEMA–3 ^b	88.32	(12.64)	84.04	(12.74)	4.28	0.82	< .001	0.34

Note: All statistics are based on the analysis of five multiply imputed datasets using a three-level hierarchical linear model, which accounts for clustering of data (students clustered within schools, which are in turn clustered within pairs of schools) and controls for baseline screener score. Means presented here are the unadjusted means for both groups.

a. Computed by dividing the estimated impact by the pooled within-group standard deviation of the TEMA–3.

b. Scores are scaled with a mean of 100 and a standard deviation of 15.

Source: TEMA–3 data collected April 2009–May 2009.

Sensitivity analyses

Six sensitivity analyses were conducted to examine the robustness of the confirmatory impact estimate to analytic choices. Because 26 schools began implementing *Number Rockets* a month later than the other 50 schools, the first sensitivity analysis examined how robust the confirmatory impact estimate was to the decision to retain the 26 affected schools in the confirmatory analysis. The second sensitivity analysis examined how robust the confirmatory impact estimate was to explicit specification of the school pairs in Level 3 of the confirmatory impact model. The third examined how robust the confirmatory impact estimate was to the decision to include the baseline covariate (pretest composite screener scores) in the model. The fourth examined how robust the confirmatory impact estimate was to the missing data approach used to address missing TEMA–3 data. This analysis used casewise deletion, where the confirmatory impact of *Number Rockets* was estimated for only those students with complete TEMA–3 scores ($n = 881$). The fifth and sixth sensitivity analyses examined how robust the confirmatory impact estimate was to the inclusion of at-risk students who were assigned to tutoring groups that included students not part of the at-risk analytic sample.⁷³ The fifth analysis excludes at-risk students ($n = 24$ excluded; $n = 970$ retained) who were assigned to tutoring groups that included students who were not identified as at risk ($n = 45$ not at-risk students). The sixth analysis excludes entire school pairs ($n = 9$ school pairs excluded; $n = 29$ school pairs retained) that included any tutoring groups with students who were not part of the at-risk analytic sample. Hierarchical linear modeling (HLM) specifications for all six sensitivity analyses are in appendix G; tables presenting full analytic output are in appendix M.

⁷³ All participating intervention schools were promised that at least nine students would receive *Number Rockets*. If fewer than nine students in an intervention school were identified as at risk, the nine students with the lowest scores on the screener were placed in *Number Rockets* tutoring groups. However, only students who met the screening criteria were included in the study analyses and posttested.

All sensitivity analyses resulted in statistically significant and positive impact estimates for *Number Rockets* (table 4-3). These estimates range from an effect size of 0.32 to 0.35, suggesting that the confirmatory impact estimate (effect size = 0.34) is relatively robust to the analytic decisions made before the evaluation. Note that one would not expect much change in the second sensitivity analysis due to the fact that most of the variance in outcomes is represented within school and not at the school-pair level.

Table 4-3. Summary of results for the six sensitivity analyses conducted on the impact of *Number Rockets* on mathematics achievement of grade 1 students, as measured by the TEMA–3, by assigned condition

<i>Sensitivity analysis</i>	<i>Outcome measure TEMA-3</i>							
	<i>Intervention</i>		<i>Control</i>		<i>Estimated intent-to-treat impact</i>			
	n	<i>M (SD)</i>	n	<i>M (SD)</i>	<i>Estimated impact</i>	<i>Standard error</i>	p	<i>Effect size^a</i>
1. Excluding affected schools ^b	414	89.19 (12.42)	261	85.35 (11.47)	3.84	(0.94)	< .001	0.32
2. Without matched pairs ^c	615	88.35 (12.64)	379	84.07 (12.74)	4.28	(0.92)	< .001	0.34
3. Without baseline ^d covariate	615	88.38 (12.64)	379	84.22 (12.74)	4.16	(0.87)	< .001	0.33
4. Using casewise deletion ^e	555	88.44 (11.45)	326	84.62 (10.29)	3.82	(0.70)	< .001	0.35
5. Excluding students assigned to tutoring groups with students who were not at risk ^b	591	88.07 (12.77)	379	83.92 (12.74)	4.15	(0.85)	< .001	0.33
6. Excluding school pairs that included tutoring groups with students who were not at risk ^b	579	88.29 (12.71)	304	84.21 (11.98)	4.08	(0.87)	< .001	0.33

Note: *M* is the unadjusted mean; *SD* is standard deviation.

- Computed by dividing the estimated impact by the pooled within-group standard deviation of the TEMA–3.
- Based on the analysis of five multiply imputed datasets using a three-level hierarchical linear model, which accounts for clustering of data (students clustered within schools, which in turn are clustered within pairs of schools) and controls for the baseline screener score.
- Based on the analysis of five multiply imputed datasets using a two-level hierarchical linear model, which accounts for clustering of students within schools and controls for the baseline screener score.
- Based on the analysis of five multiply imputed datasets using a three-level hierarchical linear model, which accounts for clustering of data (students clustered within schools, which are in turn clustered within pairs of schools).
- Based on using a three-level hierarchical linear model, which accounts for clustering of data (students clustered within schools, which are in turn clustered within pairs of schools) and controls for the baseline screener score.

Source: TEMA–3 data collected April 2009–May 2009.

Chapter 5: Exploratory analyses findings

This chapter presents the results of this study's three exploratory analyses. The first examines whether *Number Rockets* had a differential impact on grade 1 students at risk for mathematics difficulties based on baseline mathematics proficiency, as assessed by the screener. The second examines whether students who missed regular classroom instruction to attend *Number Rockets* tutoring sessions (students in the intervention condition) scored lower on a measure of word reading skill at posttest than those who did not (students in the control condition). The third examines whether the impacts of the *Number Rockets* program vary significantly depending on the average number of lessons delivered, as measured by the average number of sessions, and the Test of Early Mathematics Ability–Third Edition (TEMA–3; Ginsburg and Baroody 2003) performance (see chapter 4).

When interpreting these results, note that this study was designed and statistically powered to answer the confirmatory research question, not these exploratory questions. Thus, results reported here should be interpreted with more caution than the confirmatory impact estimate.

Exploratory analysis 1: effect of *Number Rockets* based on baseline mathematics proficiency for students participating in *Number Rockets*

A recurring issue in universal screening is the over-identification of students as being at risk (Gersten et al. 2009; Compton et al. 2010; Vaughn, Linan-Thompson, and Hickman 2003), often resulting in students receiving unnecessary services. Previous research in early mathematics screening has employed a variety of cutscores and identification procedures (for example, Mazzocco and Myers 2003). The Fuchs et al. (2005) study identified 21 percent of participating students as at risk based on a two-stage screening method; however, the screener used in the current study was not the same as the one used in the Fuchs et al. study, and no definitive guidance existed for what cutscore might be most appropriate for use with the current study's screener.

For many districts and schools intending to implement *Number Rockets*, resource constraints would be considered. For example, districts and schools might first need to identify the resources available; then some criteria (such as a cutscore on a screening test) would need to be determined to approximate the number of students who can be served by these resources.

Knowing that districts and schools will have to establish criteria to define which at-risk students will be enrolled in the intervention leads to an important question about the positive findings for *Number Rockets* reported in chapter 4—given that cutscores other than the 35th percentile could have been selected, how did the effectiveness of

Number Rockets differ depending on baseline proficiency in mathematics (as assessed by the current study's screener)?

In addition, knowing whether there is a differential impact depending on baseline mathematics proficiency might inform future strategies for school- and district-level implementation. More specifically, if *Number Rockets* has a differential effect on students by baseline mathematics proficiency, it might suggest that schools need to devote more resources⁷⁴ to universal screening to improve the precision of identifying students at risk for mathematics difficulties. Conversely, if no relationship between baseline mathematics proficiency and effectiveness of *Number Rockets* is observed, it may suggest that variability in the cutscore used for *Number Rockets* is acceptable. For example, the lowest achieving 20 students could be selected in each school, even though overall mathematics proficiency may vary across schools.

Exploratory research question 1

Does *Number Rockets* have a differential impact on grade 1 students at risk in mathematics, based on baseline mathematics proficiency?

Results indicate that there was no statistically significant interaction between baseline mathematics proficiency and the impact of *Number Rockets* (effect size = 0.08, $p = .564$). Therefore, *Number Rockets* had no statistically significant differential effect on TEMA-3 scores by baseline mathematics proficiency for the sample of at-risk grade 1 students participating in this study.

Exploratory research question 1 sensitivity analysis

In the exploratory question posed above, an analytic decision was made to treat baseline mathematics proficiency as a continuous variable. Alternative approaches were possible, such as evaluating effect sizes for students grouped by ability level. A sensitivity analysis was conducted to determine whether another analytic choice (ability-level grouping) would have resulted in statistically significant differential impact estimates. Instead of using screener scores as a continuous variable, the effect sizes for three student groups based on baseline mathematics proficiency level were compared by dividing the analytic sample into thirds at the 33rd and 66th screener score percentiles.

Despite the reduced sample size due to splitting the analytic sample into thirds ($n = 331$, $n = 331$, $n = 332$), effect sizes for the three student groups were all statistically significant: lower third (effect size = 0.34, $p = .002$), middle third (effect size = 0.31, $p = 0.012$), and upper third (effect size = 0.48, $p < .001$). (See tables M-11 through M-13 in appendix M for full results.) Therefore, regardless of grouping applied in this study,

⁷⁴ To improve the precision of screening, a longer screening battery could be used, which would require more student and teacher time to administer and score. Other alternatives include using a two-stage screening process similar to that used in the Fuchs et al. (2005) study, as described in chapter 1.

based on baseline proficiency of students at risk for mathematics difficulties, all groups were found to have statistically significant positive effects from *Number Rockets*.

An analysis comparing the treatment effect across the groups showed that they were not significantly different from each other: lower third vs. middle third (effect size = 0.07, $p = .653$); middle third vs. upper third (effect size = 0.15, $p = .373$); lower third vs. upper third (effect size = 0.08, $p = .611$).⁷⁵ (See tables M-14 and M-15 in appendix M for full results.) These results are consistent with the result for the main analysis of exploratory question 1, where student baseline mathematics proficiency was treated as a continuous variable. In both cases, there was no statistically significant relationship between student baseline mathematics proficiency and *Number Rockets* impact.

Conclusions

Overall the results from exploratory analysis 1 and the subsequent sensitivity analysis suggest that the positive confirmatory impact results may not have been different had minor changes been made to the cutscore used in this study. In other words, the confirmatory impact estimate for *Number Rockets* did not appear highly sensitive to baseline mathematics proficiency of at-risk students participating in the intervention. Because this study was designed and specifically powered for the confirmatory analysis, and not this exploratory analysis or sensitivity analysis, these results should be interpreted with caution.

Exploratory analysis 2: effect on letter- and word-reading proficiency for students participating in *Number Rockets*

Number Rockets is a supplemental pull-out program, where students are removed from their classroom during the regular school day. When students are participating in *Number Rockets*, they miss the regular classroom instruction taking place at that time. Therefore, *Number Rockets* is a tradeoff for participating students; they potentially benefit from additional mathematics instruction but lose instruction in another content area. As the emphasis on early intervention in mathematics grows (Gersten et al. 2009), a natural concern may be whether time invested in supplemental mathematics programs may reduce critical instructional time in other content areas, and students' subsequent achievement in those areas.

As described in chapter 3, the classroom instruction checklist was used to record a one-week sample of instructional activities missed by students while they were participating in *Number Rockets* (see table 3-5). Whole-class reading instruction was missed by *Number Rockets* students 8.8 percent of the time. When the various reading

⁷⁵ Note that the effect sizes reported here represent the difference in the independent effect size values of the two groups, converted to effect-size units.

activities recorded on the classroom instruction checklist were combined, they accounted for up to 33.8 percent of the classroom activities missed by students.⁷⁶ So participating in *Number Rockets* could have reduced the amount of classroom reading instruction received and, consequently, affected reading achievement.

This issue was not a primary focus of the current study, so only a single brief measure, the Woodcock-Johnson—Third Edition Letter/Word (WJ-III Letter/Word; Woodcock, McGrew, and Mather 2001) subtest, was administered at posttest immediately after the primary posttest measure, the TEMA-3. The WJ-III Letter/Word subtest assesses the ability to read letters and words aloud from a page. Because it measures only one skill important for reading, it may not be sensitive to all the possible effects of the instructional tradeoff.

Exploratory research question 2

Do grade 1 students who participate in *Number Rockets* score differently than control students on the WJ-III Letter/Word subtest?

There was no statistically significant relationship between participation in *Number Rockets* and performance on the WJ-III Letter/Word subtest (effect size = -0.01 , $p = .913$). Note that the current study was not explicitly designed or statistically powered to look for such adverse instructional tradeoffs. Also, the WJ-III Letter/Word measure does not assess the breadth of reading skills taught in the regular classroom; it may not have been sensitive to adverse impacts of student acquisition of other reading skills.

Exploratory analysis 3: Relationship between program impacts and average number of delivered lessons

This study targeted delivery of a minimum of 45 *Number Rockets* tutoring sessions to each tutoring group. However, the actual school-average number of tutoring sessions delivered to each student group ranged from 37.3 to 56.1 across intervention schools. Because of this variability in session delivery and the fact that this study employed a blocking design where schools were paired prior to randomization, it is possible to examine dosage issues to determine if the observed intervention effect on student TEMA-3 performance (see chapter 4) varied based on the school-average number of tutoring sessions delivered to tutoring groups at intervention schools. In this exploratory analysis, dosage is the average number of sessions delivered per tutoring group in each intervention school.⁷⁷ A variance decomposition analysis indicates that

⁷⁶ Since more than one type of reading instruction could be missed while a student attended a single *Number Rockets* lesson, this percentage reflects an upper limit.

⁷⁷ The variability in the number of tutoring group sessions delivered *within* schools was compared with the variability in the number of tutoring group sessions delivered *between* schools. The level 2 (school) intraclass correlation (ICC) was 0.199 and the level 1 (groups within schools) ICC was 0.801. Therefore,

most of the variance in lessons delivered to student groups was observed within schools as opposed to between schools. Only between-school variability was relevant to this research question, thus substantially reducing its statistical power to detect potential dose-impact relationships at the school-pair level.

Knowing that districts and schools will have to make decisions concerning how much of the intervention to implement (based on resources such as time and personnel), naturally leads to an important question on the relationship between the level of implementation of the intervention and the effect of that intervention.

Exploratory research question 3

Do the impacts of the Number Rockets program vary significantly depending on the average number of lessons delivered?

There was no statistically significant relationship between the school-average number of *Number Rockets* sessions delivered to *Number Rockets* tutoring groups in each intervention school and the intervention effect on student TEMA-3 performance (effect size = 0.06, $p = .576$). Therefore, higher levels of implementation of *Number Rockets* were not associated with larger impacts on TEMA-3 performance in this study.

In this exploratory analysis, dosage is limited to the range of average sessions delivered to schools in the present study; this study was not specifically designed or powered for this exploratory research question, as most of the variability in sessions delivered was observed within-schools as opposed to between schools. Also, the number of sessions was not randomly assigned to the school pairs and, therefore, this analysis is correlational.

some of the variability in number of sessions delivered was attributable to school context and some of the variability is due to other factors (such as individual tutors or scheduling). However, given the study's design, school-level random assignment from matched pairs was used to estimate dose-impact relationships.

Chapter 6: Summary of key findings and study limitations

This chapter summarizes the findings for the first effectiveness study of *Number Rockets* as a Tier 2 mathematics intervention, reviews the results of the fidelity of implementation data, discusses important limitations to be considered when interpreting the results, and suggests directions for future research.

Effect of *Number Rockets* on mathematics achievement

The main finding of this effectiveness study is that students at risk for difficulties in grade 1 mathematics benefited by participating in *Number Rockets*. Participation had a statistically significant difference (effect size = 0.34) on Test of Early Mathematics Ability—Third Edition (TEMA–3; Ginsburg and Baroody 2003) scores favoring the intervention group over the control group ($p < .001$). This finding was observed in a sample of students from four urban districts across four states. In addition, this finding is robust when applying alternate analytic strategies such as the exclusion of the 26 schools affected by a natural disaster and the specification of matched school pairs in the hierarchical linear model.

The current study's effect size of 0.34 standard deviations is smaller than the effect sizes for all four outcome measures demonstrating statistically significant results in the Fuchs et al. (2005) study (statistically significant effect sizes ranged from 0.40–0.70). This was expected given the current study's emphasis on implementing *Number Rockets* in conditions more closely resembling what urban school districts experience in their day-to-day instructional environment when implementing interventions. The observed lower levels of fidelity of implementation (see chapter 3) are consistent with this expectation.

Several other factors may have contributed to the smaller effect size. First, students participating in *Number Rockets* missed some portion of their core mathematics curriculum during 11.4 percent of the lessons sampled study-wide (see table 3-5). Despite the intention that *Number Rockets* serve as a supplemental intervention, it replaced a portion of regular mathematics instruction for some students in the study.⁷⁸ Second, while the constructs intended to be assessed by both sets of outcome measures were similar (the TEMA–3 in the current study and the seven measures of mathematics skills in the Fuchs et al. [2005] study: Story Problems, First-Grade Concepts/Applications, WJ–III—Calculation, Curriculum-Based Measurement—Computation, Addition Fact Fluency, Subtraction Fact Fluency, and WJ–III—Applied problems), the assessment contexts differed. The TEMA–3 was administered by a single staff member, and nearly all items

⁷⁸ The Fuchs et al. (2005) study reported that *Number Rockets* students missed an average of 10.56 minutes of regular mathematics instruction during the intervention.

were untimed. In the Fuchs et al. study, three of the four statistically significant outcome measures were group administered to an entire classroom, and three of the four were explicitly timed. Third, in the Fuchs et al. study, effect sizes were calculated as change from pretest to posttest on the same measures and test forms. The use of identical measures for pretest and posttest may increase sensitivity to change, if it truly occurred on the type of knowledge and skills assessed. In the current study, the pretest screener was a set of six subtests, different from the TEMA–3 outcome measure. Fourth, the Tier 1 instructional context around *Number Rockets* also differed between the two studies. To prevent contamination, this study restricted tutors from communicating with classroom teachers about student progress. In the Fuchs et al. study, however, teachers were given regular updates on the progress of their students in tutoring. They were also provided with teaching strategies by coaches, with the intention that teachers would use the information to improve whole-class instruction and provide differentiated instruction to at-risk students. Finally, the Fuchs et al. study district used a single curriculum; each of the four urban districts in the present study used a different core mathematics curriculum, which may have provided a more heterogeneous instructional context.⁷⁹

Effect of *Number Rockets* by baseline mathematics proficiency of at-risk students

As reported in chapter 5, the study team did not find a statistically significant relationship between students' baseline mathematics proficiency and the impact of *Number Rockets*. Due to the variety of screening strategies in the literature (for example, Gersten, Jordan, and Flojo 2005; Fuchs et al. 2007), there is currently no established screening tool or fixed proportion of students who might be identified as at risk. Therefore, districts and schools will most likely determine their own strategies for doing so.

The finding that student proficiency at baseline was not related to *Number Rockets* outcomes suggests that a range of cutscores or student selection strategies might be successful in identifying students who might benefit from *Number Rockets*. An effective cutscore depends on a number of factors, including the screening measure used, alignment of the content of the screener with that of the intervention, and the particular sample from which the cutscore is determined. Therefore, caution should be used in adopting a particular means of determining at-risk status developed in another context, such as in this study. Note also that the current study was not explicitly designed or statistically powered to compare the decision accuracy of various cutscores. This is an empirical question to be evaluated in future studies.

⁷⁹ Each participating district used a different core mathematics curriculum: enVision Math™, Houghton Mifflin Math™, Math Investigations™, or Scott Foresman-Addison Wesley™.

Effect of participation in *Number Rockets* on reading achievement

As reported in chapter 5, there was no statistically significant relationship between students' participation in *Number Rockets* and outcomes on the Woodcock-Johnson—Third Edition Letter/Word (WJ-III Letter/Word) subtest. This subtest was administered to evaluate whether removal from the regular classroom, and any concomitant reduction in reading instruction, may have adversely affected students' letter- and word-reading skills. Of the sampled *Number Rockets* lessons, 8.8 percent resulted in students missing whole-class reading instruction.

Despite these substantial percentages, there was no evidence of adverse impact of *Number Rockets* participation on letter- and word-reading skills from the exploratory analysis reported in chapter 5. Note that the current study was not explicitly designed or statistically powered to look for such adverse instructional tradeoffs. Further, the WJ-III Letter/Word measure used here is a focused measure of word-reading skills and does not assess the breadth of reading skills taught in the regular classroom.

Effect of *Number Rockets* by the school-average number of tutoring sessions

There was no statistically significant relationship between the school-average number of *Number Rockets* sessions delivered and the intervention effect on student TEMA-3 performance (effect size = 0.07, $p = .667$). This exploratory analysis was not sensitive enough to rule out that a dosage-impact relationship at the school-pair level might exist. The study was not specifically designed or powered for this exploratory research question, and this exploratory finding is limited to the range of school-average tutoring sessions delivered in the present study.

Implementation of *Number Rockets* in a real-world context

As an effectiveness trial, this study departs in a number of important ways from the Fuchs et al. (2005) efficacy study, in terms of implementation. First, the personal background characteristics of the tutor pool in the current study was more heterogeneous than that in the Fuchs et al. study. In the current study, over 66 percent of the tutors held a teaching certificate, 44 percent had five or fewer years of teaching experience, and 77.3 percent were retired and/or substitute teachers. The tutors in the current study represented the types of individuals typically available for hire by participating districts. In the Fuchs et al. study, 83.3 percent of the tutors were Master's-level graduate students in special education or school counseling, all tutors had worked closely with the *Number Rockets* developer, and none had a teaching certificate. Second, training in this study was typical

of the level of professional development provided by publishers of instructional curricula to district personnel: one full day of training, followed by two two-hour sessions and telephone and email support (Agodini et al. 2009). This was less tutor support than was provided in the Fuchs et al. study, which held weekly meetings and conducted tutor observations followed by active coaching of tutors throughout implementation. Finally, the number of lessons delivered in the present study averaged approximately 48.37 per tutoring group, with substantial variability ($SD = 6.45$), and average lesson fidelity of 85.0 percent. The Fuchs et al. study reported that 48 lessons were delivered to each group and average lesson fidelity was 94.6 percent.

Study limitations and future research

This study is the first effectiveness evaluation of *Number Rockets* and builds on the positive findings of the Fuchs et al. (2005) efficacy study. An effectiveness study was a next step in continuing to establish an evidence base for this intervention; however, any single study has limitations that must be considered when interpreting the findings.

First, the control (or counterfactual) condition was the absence of additional mathematics instruction for the at-risk students beyond regular classroom instruction. Therefore, it cannot be stated whether the effects described in chapter 4 are due to additional mathematics instruction time delivered in any manner or to specific characteristics of *Number Rockets*. In part, this issue results from a lack of Tier 2 interventions suitable for large-scale evaluation to provide a reasonable counterfactual to *Number Rockets*. It would, therefore, be important to compare *Number Rockets* with a counterfactual condition that controlled for additional instructional time in mathematics, either through adding time with the existing classroom mathematics curriculum or through using a different Tier 2 intervention to supplement regular instruction. This would distinguish between effects from additional instructional time in a generic sense, versus effects due to the instructional delivery model defined by *Number Rockets*.

Second, the requirement of active parent consent in this study introduced a potential for student selection bias after random assignment of schools. Differential consent form return rates were observed between the intervention and control schools. No data were collected allowing the study team to evaluate the effort of school personnel to collect consent forms, so it cannot be known whether the differential rate observed was due to school personnel in intervention schools expending more effort to achieve high consent return rates, to parent awareness of the assignment status of their child's school, or to some other reason. This differential rate may have introduced bias between the intervention and control groups at baseline on unobserved variables

Third, specific urban districts were recruited for this study, and the students included represent a sample whose parents gave them consent to participate. Because districts and schools volunteered for the study, they are not statistically representative of

a larger population. Therefore, the main finding applies only to these districts, schools, and students. Ideally, a nationally representative sample of districts and schools would be recruited in future evaluations so that findings could be generalized beyond a sample.

Fourth, *Number Rockets* is not currently available in Spanish. In study districts, English-language learner students comprised from 1 percent to 29 percent of students across all grades (National Center for Education Statistics; n.d.). Therefore, as a Tier 2 intervention for grade 1, *Number Rockets* did not address the needs of the at-risk students who received mathematics instruction in Spanish in the study schools, nor would it address the needs of these students in many other schools.

Fifth, the impact estimates focus on the outcomes at the end of grade 1. The study does not provide evidence on the persistence of the benefits of *Number Rockets* participation. It is not known if students who benefited in grade 1 would be better prepared for success in mathematics at the beginning of grade 2 or beyond. Future research questions could focus on whether *Number Rockets* students are just as prepared for grade 2 after an intervening summer with no Tier 2 instruction, or if they are less likely to require future Tier 3 instruction or special education referrals if they do participate in Tier 2 instruction.

Sixth, the use of a study-wide cutscore for determining at-risk status might not reflect the approach of some local education agencies for universal screening. For example, some local education agencies may allow each campus to rank-order students at the campus-level and classify a fixed number or percentage of the lowest performing students as eligible for a Tier 2 intervention. Other methods of identifying at-risk students may result in an at-risk student sample with different ability levels. Furthermore, the study was not designed to provide information about specific grade 1 mathematics skills (such as understanding place value) that may be indicators of at-risk status.

Seventh, that tutors were required to record the audio of each delivered lesson might be an additional limitation to the study's generalizability. That is, because tutoring sessions were being monitored, tutors may have implemented *Number Rockets* with greater fidelity than they would have had the tutoring sessions not been monitored.

Eighth, tutor turnover was not trivial. Approximately 34% of student groups experienced a change in tutor, and approximately 9% experienced more than one change in tutor. It is not known how representative this turnover rate would be of turnover in other districts attempting to implement *Number Rockets*. It is possible that the use of a temporary employment agency affected turnover in ways different from what districts hiring tutors directly may experience. An exploratory analysis of any possible relationship between tutor turnover and student outcomes could not be conducted for this report.

Finally, tutors were instructed not to communicate with classroom teachers about individual student performance. This constraint was imposed to prevent teachers from using information from *Number Rockets* in an ad hoc manner, which could have introduced contamination of strategies into the classroom. The constraint might be relaxed in a real-world implementation of a Tier 2 intervention.

Further studies could examine the tradeoffs for the level of tutor professional development and tutoring provided, such as evaluating whether 20–30 minute intervention sessions would be as effective as the 40-minute intervention sessions of the current study. Future studies could also be designed and powered to follow up on the exploratory finding that, within the range of school-average tutoring sessions delivered in this study, there was no statistically significant relationship between the average number of sessions and school-pair impact estimates for *Number Rockets*. This follow-up could be done to determine, for example, if a minimum number of sessions were required to achieve the impact observed in the present study.

Appendix A: Study timeline

Table A-1. Dates of study phases by district

<i>Study phase</i>	<i>Schools not affected by natural disaster</i>	<i>Schools affected by natural disaster</i>
Recruit schools	October 31, 2007–May 30, 2008	October 31, 2007–May 30, 2008
Conduct random assignment	June 30, 2008–July 31, 2008	June 30, 2008–July 31, 2008
Distribute/collect parent consent forms	August 27, 2008–October 24, 2008	August 27, 2008–December 18, 2008
Train screener assessment teams	October 13–30, 2008	October 27–28, 2008
Screen students	October 13, 2008–November 7, 2008	October 27, 2008–December 19, 2008
Train tutors	November 13–19, 2008	November 10–11, 2008; and January 5, 2009
First follow-up training	January 15 and 20, 2009	January 29, 2009
Second follow-up training	February 17 and 19, 2009	March 19, 2009
Implement <i>Number Rockets</i>	December 3, 2008–May 8, 2009	January 29, 2009–May 15, 2009
Train post-test assessment teams	April 23, 2009– May 8, 2009	May 7–8, 2009
Collect posttest data	April 27, 2009–May 27, 2009	May 11–12, 2009

Source: Study team records collected October 2007–May 2009.

Appendix B: Power analysis assumptions

Minimum detectable effect size

The minimum detectable effect size (MDES) represents the smallest *true* program impacts in standard deviation units that can be detected with high probability (Bloom 2005). All other things being equal, the smaller the effect size to be detected, the larger the study sample must be. The MDES selected should be large enough that the impact is an important one to detect but small enough to be feasible given the intervention. In a randomized efficacy trial of the *Number Rockets* intervention, Fuchs et al. (2005) found statistically significant effect sizes ranging from 0.40 to 0.70 for four mathematics screening measures for the *Number Rockets* intervention group. In a synthesis of meta-analyses of educational interventions, Hill et al. (2008) reported a mean effect size of 0.23 for interventions for grades 1–3. The Institute of Education Sciences (2011) considers an effect size of 0.25 to be a meaningful finding even if nonsignificant due to a lack of statistical power.

The appropriate MDES for this study was informed by these estimates. This study was powered for an MDES of 0.27 under the most conservative assumptions in table B-1 regarding intraclass correlations (ICC) and R^2 for the baseline covariate. Initially targeted was an MDES of 0.30, which preliminary power calculations had indicated would require approximately 60 schools. However, due to the opportunity provided by a strong response to early recruiting and given the likely increased school attrition risk when conducting an effectiveness trial, the study team increased the sample target to 70 schools.

While relatively conservative given the results of the Fuchs et al. (2005) study, the MDES of 0.27 incorporates the strong prior evidence supporting this intervention while recognizing the study's design as an effectiveness trial, and it is still large enough to represent important education gains, especially when averaged across a large number of students. The following power analysis tables provide the MDES for various sample sizes and assumptions.

Assumptions made for power analysis

The power analysis assumed a design in which schools were randomly assigned to a condition, with approximately three grade 1 classrooms per school. The power analysis assumed a design to model clustering at the school level only. The power calculations were based on the following additional assumptions:

- Desired statistical power—80 percent.
- Statistical significance level— $\alpha=0.05$ for a two-tailed test.

- Number of students per classroom—Assumed that each classroom had 20 students with 20–25 percent of the students Tier 2 eligible; a 20 percent attrition rate yielded 3–4 students per classroom and an average of 9–12 students per school.
- Proportion of schools in the intervention condition—50 percent employing matched pairs (Ginsburg and Baroody 2003).
- ICC—Assumed an ICC of 0.13 for both the school and classroom levels. (Results are also provided for values of 0.10 and 0.15.)
- Explanatory power of the covariates—Assumed that the covariates (pretest scores on the screening battery) would yield a value of $R^2 = 0.40$ for the Test of Early Mathematics Ability–Third Edition, with resultant error reduction for both the school and classroom levels. Results are also provided for values from 0.30 to 0.70; no additional covariates were used.

Basis for the assumptions

The assumptions for this study were primarily conventional, and there were no compelling reasons to modify them. These included a power level of 0.80 and setting α at 0.05.

The screening criterion was selected to be a cutscore representing the 35th percentile on the screener, a criterion used in several other studies to determine at-risk status in early mathematics (Hanich et al. 2001; Jordan, Kaplan, and Hanich 2002; Jordan, Hanich, and Kaplan 2003). The 35th percentile also ensured that at least one at-risk student was identified in each study school. Selecting different cutscores may result in different numbers of at-risk students identified (and therefore clustered) within schools, influencing power. Schochet (2005) provides tables of school-level ICCs for various measures used as outcome measures in a range of studies and recommends using the same ICC values for the classroom level. These ICCs for early elementary grades (1–3) in mathematics ranged from 0.03 to 0.19 when using standardized test scores, with an average of 0.13. The use of ICC = 0.15 appeared to be conservative for this study, given the ICCs from other interventions as cited in Schochet (2005). However, since there are no specific estimates for these interventions, the power analysis also examined ICC values of 0.10 and 0.15.

As noted above, a single baseline covariate was used in the analyses and it was assumed that the amount of variance in the outcomes explained by the covariate would be at least 0.50 (in effect, $r_{\text{pre-post}} \approx 0.71$). This assumption is most likely conservative. (See Bloom, Richburg-Hayes, and Black [2007] for R^2 estimates for school-level covariates.) For example, Baker et al. (2006) found a correlation of 0.72 for the Number Knowledge Test (Baker et al. 2006) as a pretest measure and the Stanford Achievement Test Series,

Ninth Edition (Harcourt Assessment, Inc. 2004), as a posttest measure. But because in this study the analyses were conducted with only at-risk students, a restriction of range effect could have reduced the corresponding R^2 values by an unknown amount. Therefore, a lower R^2 value of 0.40 was assumed, though a range of R^2 values from 0.30 to 0.70 were examined.

Target sample size

A target sample size of a minimum of 70 schools (35 intervention and 35 control) was selected because it provided sufficient power (for an MDES of 0.30 or better) even under the conservative assumptions outlined above. It would also provide some insurance against school attrition. Table B-1 includes findings from the power analysis for 70 schools incorporating the assumptions above.

Table B-1. Power analysis (minimum detectable effect size for minimum of 70 schools)

<i>Intraclass correlation</i>	<i>Number of schools (intervention + control)^a</i>	R^2				
		<i>0.30</i>	<i>0.40</i>	<i>0.50</i>	<i>0.60</i>	<i>0.70</i>
0.15	70	0.27	0.25	0.23	0.20	0.17
0.13	70	0.26	0.24	0.22	0.19	0.17
0.10	70	0.24	0.22	0.20	0.18	0.15

a. Assuming matched pairs.

Source: Authors' calculations March 2008.

Even under more conservative assumptions, there would still be sufficient power to obtain the MDES in the desired range. The MDES remained in the range of 0.24 to 0.27, even when the ICCs were as low as 0.10 or as high as 0.15⁸⁰ (keeping other assumptions constant) or if the R^2 was as low as 0.30 (again, keeping other assumed values constant).

Due to schools' continuing interest in participating and the need to include all eligible schools in some districts as a condition of those districts' participation, 82 schools were ultimately enrolled in the study before school random assignment; subsequent analytical issues and the attrition of one school pair reduced the number of eligible schools retained for the final analysis to 76 (38 pairs).

⁸⁰ Before estimating the confirmatory hierarchical linear impact model, an unconditional model (without treatment status or any covariates) was estimated to evaluate the proportion of variance accounted for by clustering at the school and pair level. The estimated intraclass correlation between schools was 0.059 and between pairs was 0.006.

Exploratory analyses

The power analysis determined the target sample size for the primary confirmatory analysis. Using the target sample size of 70 schools, the MDES for the exploratory analyses can also be calculated.

For exploratory analysis 1, the assumptions from the power analysis for the primary confirmatory analysis remain appropriate, and the $MDES = 0.24$. For exploratory analysis 2, the model was the same as for the primary confirmatory analysis, except for a different outcome measure. The assumptions used for the power analysis are therefore appropriate, except for the explanatory power of the pretest covariate, assumed to be $R^2 = 0.10$. This is approximately equal to a correlation of 0.32 between the current study's screener and the Woodcock-Johnson—Third Edition Letter/Word subtest (Woodcock, McGrew, and Mather 2001). This yields an MDES of 0.29 for exploratory analysis 2. For exploratory analysis 3, the assumptions for the power analysis are the same as for the primary confirmatory analysis.

Appendix C: Parent consent form

PARENT/GUARDIAN CONSENT FORM

August 2008

Dear Parent or Guardian,

We are delighted to announce that your child's school and district have agreed to be a partner in a ground-breaking study. We are contacting you to provide information about the research study taking place in your child's school during the 2008-2009 academic year, and asking for you and your child's voluntary participation. This study will include all first grade classrooms and students. The study examines the use of an intervention, *Intensive Small Group Mathematics Instruction*, for first grade students who have been identified as *at-risk* in mathematics. It is funded by the U.S. Department of Education and conducted by the Regional Educational Laboratory – Southwest (REL Southwest). Education research is critical in providing guidance for programs that will improve student achievement. In this study we hope to learn how to help students who may be *at-risk* for falling behind grade-level in mathematics.

All first grade classrooms and students at your child's school will be invited to participate in this study. In your district, participating schools will be randomly assigned as an *intervention* school (*at-risk* students identified through screening will receive tutoring in mathematics) or a *control* school (will not receive tutoring). In all participating schools, all students in first grade will complete two sets of mathematics tests; one in fall and one in spring. In the *intervention* schools, students who have been identified as *at-risk* in mathematics, based on the fall math test, will receive 17-weeks of additional math instruction in small group tutoring sessions. Students leave their regular classroom to work in small groups with a trained math interventionist (tutor). These types of programs are currently common for reading in schools nationwide. For qualifying *at-risk* students, the math tutoring sessions will begin in December and end in April.

How will my child participate in this study?

With your permission, your child will be administered a *mathematics screener* once in the fall and a general math achievement measure at the end of the school year. The tests are similar to typical mathematics tests given in the school district. The test will be administered at your child's school, in cooperation with school officials and teachers. Testing time will be scheduled with your child's teacher so that your child does not miss class work. All students will receive the district's regular mathematics instruction.

If your child's school is randomly assigned to receive the intervention, scores on the mathematics screener will determine whether your child qualifies to receive 17 weeks of additional math instruction. Tutoring is provided in small-groups of 2 or 3 students, working with a trained math tutor in 40 minute sessions. These groups will meet during the school day for approximately 50 sessions; two to three times per week between December 2008 and May 2009. Tutoring sessions will be audio-taped to ensure the quality of the lesson.

What are the risks and benefits for my child to participate in this research?

The research conducted in your child's school district is designed to determine how to help students who may be *at-risk* for falling behind grade-level in mathematics. *If* your child's school is selected at random to be an *intervention* school, *and if* your child identified as being *at-risk* in mathematics through screening, your child would receive over thirty-hours of focused math tutoring between December 2008 and May 2009 during the school day. **All tutors will have previous elementary teaching experience, and most (if not all) will be certified teachers. In addition they will be closely monitored throughout the study to ensure high quality instruction is provided to your child.**

Previous research indicates that this intervention has a significant positive impact on student mathematics achievement for *at-risk* students. All students in the district will continue to receive the district's normal mathematics related instruction.

There are *no* physical risks or discomforts of any kind. You should however be aware that tutoring is provided during regular school hours. Your child will not be in the regular classroom during the time that tutoring is provided. Some regular classroom time will be missed. The study team will work closely with teachers and school personnel to minimize impact on core academic instruction, such as for math and reading.

Will information about my child remain private?

We will not share the information collected from or about your child with anyone outside the research team. Your child's test results will only be used to evaluate the effectiveness of the mathematics intervention. **The reports prepared for this study will summarize findings across the students, classrooms, and schools and will *not* associate information with a specific child.** Responses for this data collection will be used only for statistical purposes. We will not provide information that identifies you, your child, your child's school or district to anyone outside the study team, except as required by law.

No child's name will be used or appear in any written work. In addition, your child's participation in the study will not affect his or her treatment at school. Please note that you may withdraw your child from assessment or small-group instruction at any time without penalty. Your child also does not have to answer any test questions he or she does not want to answer.

We hope you will see the value of this research and agree to consent to your child's participation. Please sign below and return this letter to your child's teacher.

If you have questions about this study, please contact Denise Clyburn, the Project Coordinator by email at dclyburn@edvanceresearch.com or by telephone toll-free at 1-877-338-2623, extension 4115 between the hours of 8am and 5pm Central Standard Time. If you have concerns or questions about your child's rights as a participant, contact the Chair of AIR's Institutional Review Board (which is responsible for the protection of study participants) at IRBChair@air.org, toll free at 1-800-634-0797 or c/o AIR, 1000 Thomas Jefferson Street, NW, Washington, DC 20007.

Sincerely,

Dan Hunt
Project Manager

First-Grade Mathematics Tutoring Study (Math-RTI) **Parent/Guardian Consent Form**

Please check the appropriate box and fill in the information below. Please return this form to your child's teacher. **THANK YOU!**

☐ **YES, (1) My child MAY participate in the Math-RTI Study mathematics tests, which includes a 30 minute math screener administered in the Fall, and 90-150 minute math assessment given in the Spring.**

(2) I also agree *that if* my child attends a school that is randomly assigned to receive the intervention; *and if* my child is identified as *at-risk* based on the Fall screening test; my child MAY receive 17-weeks of small group instruction with a trained math tutor and 1 or 2 other children.

I understand there will be approximately 50 sessions, each lasting 40 minutes; 2-3 sessions per week for 17 weeks. These will be conducted approximately from December 2008–May 2009.

I understand that during the time my child participates in tutoring sessions, my child will not receive instruction in their regular classroom, and may miss some instructional time on topics other than math. The study team will work closely with the school personnel to minimize any impact.

I understand that regular math instruction will be provided to all students by the regular classroom teacher just as before; and no classroom time will ever be taken from regular math instruction for purposes of additional tutoring.

(3) I agree (if my child does receive tutoring as part of this study), to allow my child's group tutoring sessions to be audio-taped for the purposes of quality control and monitoring of the tutor, and for no other purpose.

(4) I understand that my child may withdraw from the study at any time and for any reason with no penalty whatsoever.

☐ **NO, I DO NOT want my child to participate in Math-RTI Study-related mathematics tests, or any portion of the study.**

I understand that the specific math screening and tutoring activities provided to the district *as part of this study* will not be available to my child at a later date during the current academic year.

I have read the above information. I have asked any questions I may have and received answers. My response is indicated in one of the check-boxes above.

Parent or Legal Guardian Name *(please print)*

Child's Name *(please print)*

Parent or Legal Guardian Signature _____

Date _____

Appendix D: Screener subtest details and descriptive statistics

Table D-1. Screener for current study

<i>Subtest</i>	<i>Source</i>	<i>Construct</i>	<i>Activity</i>	<i>Reliability^a</i>	<i>Time limit</i>	<i>Sample item</i>
Quantity Discrimination	Clarke et al. 2006	Ability to make numerical judgments of magnitude	Student verbally identifies the greater of a visually presented pair of numbers	$\alpha = 0.93$	1 minute	11–9
Curriculum-Based Measurement–Computation	Fuchs et al. 2005	Early mathematics achievement outcomes	Student writes responses to single- and double-digit addition and subtraction problems	$\alpha = 0.95$	2 minutes	10 – 3
First Grade Concepts/ Applications	Fuchs et al. 2005	Ability to solve applied problems using grade 1 mathematics skills	Administrator reads word problems aloud. Student uses visual stimuli on a student worksheet to write responses to items such as time, shape, and length	$\alpha = 0.92$	15–30 seconds per item (Total of all items sums to 7 minutes)	A B C D E F G Write the eighth letter. _____
Number Knowledge Test	Baker et al. 2006	Number sense	Administrator reads items aloud. Student identifies the number that best matches the relationship between the numbers	$\alpha = 0.98$	Not timed; average completion time: 8 minutes	Which number is closer to 7: 4 or 9?
Story Problems	Jordan et al. 2007	Ability to solve word problems in which objects are referred to but not presented	Administrator reads problems aloud. Student verbally responds to addition and subtraction word problems	$\alpha = 0.58–0.77^b$	Not timed; average completion time: 1–2 minutes	Alex has 2 pennies. Maria gives him 4 more pennies. How many pennies does Alex have now?
Digit-Span Backward	Geary 1993	Auditory working memory, concept development, working memory, and speed of processing	Administrator reads a series of numbers aloud. Student says the numbers in reverse order from which they were read	Not provided	15 seconds per trial	5–7–4 Correct response: 4–7–5

a. Coefficient alphas, representing estimates of internal consistency reliability, are reported and drawn from the respective cited sources. Alphas related to the item-level data for the current study are not available because data were entered at the total score level.

b. The Jordan et al. (2006) study reported coefficient alphas for four administrations of this test across the academic year; this range encompasses all four coefficients.

Source: Authors' summary of citations listed in source column.

Table D-2. Descriptive statistics for the six screener subtests and the screener composite score

	<i>Mean</i>	<i>Standard deviation</i>	<i>Correlations</i>					
			<i>QD</i>	<i>CM</i>	<i>CA</i>	<i>NKT</i>	<i>SP</i>	<i>DS</i>
Quantity Discrimination (QD)	22.25	(9.88)						
CBM Computation (CM)	6.89	(4.12)	0.48					
First-Grade Concepts/Applications (CA)	12.06	(4.40)	0.58	0.60				
Number Knowledge Test (NKT)	15.71	(5.44)	0.63	0.51	0.67			
Story Problems (SP)	4.70	(2.48)	0.44	0.47	0.56	0.53		
Digit-Span Backward (DS)	2.21	(1.58)	0.44	0.38	0.46	0.48	0.39	
Screener composite ^a	-0.03	(0.79)	0.77	0.75	0.84	0.82	0.74	0.68

Note: CBM is Curriculum-Based Measurement. All correlations statistically significant at $p < .001$; results for all screened students in 76 retained schools ($n = 2,719$); $n = 2,708$ for Digit-Span Backward; $n = 2,718$ for Quantity Discrimination. Composite score calculated for those students based on remaining subtests. All subtest means and standard deviations in this table are based on subtest raw scores.

a. Original composite was formed from all screened students in 52 participating schools, including schools not retained in the study, then applied to the students in the 26 schools affected by the natural disaster. The composite mean and standard deviation reported in this table are based on averaged subtest z -scores.

Source: Authors' analysis of screener data collected October 2008–December 2008.

Appendix E: Student mobility

Table E-1. Mobility for students in the analytic sample

<i>Student location</i>		<i>Across all districts</i>
<i>At screening</i>	<i>At posttest</i>	
Intervention school	Nonparticipating district or nonparticipating school	41 ^a
Control school	Nonparticipating district or nonparticipating school	34 ^a
Intervention school	Intervention school	5
Control school	Control school	0
Crossovers		
Intervention school	Control school	3
Control school	Intervention school	0
Total		83

Note: Students who moved between study schools were posttested as members of the group to which they were originally assigned.

a. A total of 75 students moved out of study schools and districts but were still included in the analytic sample under intent-to-treat. Their Test of Early Mathematics Ability–Third Edition (TEMA–3; Ginsburg and Baroody 2003) scores were estimated. This count matches the $n = 75$ reported in chapter 2, as does the count of students missing TEMA–3 scores because of mobility.

Source: Instructional log data collected December 2008–May 2009.

Appendix F: Fidelity measures

Figure F-1. Sample lesson fidelity checklist

First Grade Math Project: Tutoring Fidelity			
Checked by: _____ Date: _____ Tutor: _____ Session #: _____ Session Date: _____			
School: _____ Teacher(s): _____ Group Size: _____			
+ = behavior observed - = behavior not observed NA = not applicable			
Note: Make notes of any variations from the steps.			
TOPIC 15 (Day 1)			
+	-	NA	Step
			1. The tutor presents review sheet #14 and reads each question aloud, allowing time for students to write answers.
			2. The tutor distributes Topic 15 Day 1 Worksheet 1 and states that they will work on a different kind of addition problem.
			3. The tutor explains that the first problem has a number in the 1s place and the 10s place.
			4. The tutor explains that they will use Base Ten Blocks to help add $85 + 12$ and that he/she first will show 85 with the blocks.
			5. The tutor explains that because there are 8 10s in 85, he/she will put 8 rods in the 10s place.
			6. The tutor explains that since there are 5 1s in 85, he/she will put 5 cubes in the 1s place.
			7. The tutor repeats steps 5 and 6 to show the number 12.
			8. The tutor counts rods and cubes again to make sure numbers are correct.
			9. The tutor states that in two-digit addition we always start in the ones place and demonstrates moving all cubes to the bottom square and recording the total number in the 1s place.
			10. The demonstrates moving all rods to the bottom square and recording the total number in the 10s place.
			11. The tutor counts the total blocks again and reads the final number sentence aloud.
			12. The tutor follows the same procedure for the remainder of the worksheet, allowing students to take turns showing the numbers with the blocks, counting the total blocks, and reading the number sentences aloud.
			13. <i>If a student does not perform tasks correctly, the tutor provides assistance.</i>
			14. <i>If time permits, tutor provides Topic 15 Day 1 Worksheet 2 and continues with same procedure.</i>
			15. Throughout the lesson, tutor uses a behavior modification method. (For example: timer, rocket ship, awarding points)
Total (+)		Total (+) and (-)	
Proceed to flash card fidelity check (separate sheet). Total fidelity score for the session is based on lesson fidelity and flash card activity fidelity.			

Source: Paulsen and Fuchs 2005.

Table F-1. Example of aggregated instructional log data

<i>Group</i>	<i>Lesson</i>	<i>Date taught</i>	<i>Tutor</i>	<i>Session time (minutes)</i>	<i>Comments</i>	<i>Lesson absentees</i>
2	Topic 12, Day 4	3/30/2009	Tutor 1	40		Student 1 Student 2
5	Topic 15, Day 1	3/30/2009	Tutor 2	38	good day	
12	Topic 12, Day 1	3/30/2009	Tutor 3	40		
15	Topic 11, Day 6	3/30/2009	Tutor 4	50	went very well	
7	Topic 11, Day 6	3/30/2009	Tutor 5	39		
26	Topic 12, Day 5	3/30/2009	Tutor 6	44		Student 3
18	Topic 12, Day 1	3/30/2009	Tutor 7	40		
21	Topic 12, Day 2	3/30/2009	Tutor 8	41	Student 5 missed the first 15 minutes	
9	Topic 12, Day 2	3/30/2009	Tutor 9	40		
4	Topic 12, Day 1	3/30/2009	Tutor 10	41	Student 6 moved	Student 7
23	Topic 11, Day 6	3/30/2009	Tutor 11	37		

Note: All data in this table are fictitious.

Source: Authors' simulation of actual session logs.

Figure F-2. Classroom instruction checklist**What classroom activities are the students in your First Grade Math Tutoring group missing?**

For the week of collection complete the following form each day for the group you are tutoring.

Complete one form each day the group meets. For example, if the group meets 3 days a week then you would complete 3 forms for this group, if the group meets 4 days a week then you would complete 4 forms, one form per day, per group.

Group Number: _____

Date: _____

School Name and District: _____

Tutor Name: _____

Indicate what activities/classes the students are missing by placing a check in the appropriate box. Check all boxes that apply. If one of the students in your group is absent or if you have less than 3 students in this group go to the next column. Remember to return the completed forms to the Kelly Services office at the end of the week.

Student Name: _____

Student Name: _____

Student Name: _____

Teacher Name: _____

Teacher Name: _____

Teacher Name: _____

Reading

- ☐ Whole class reading instruction such as phonics, phonemic awareness, fluency, vocabulary, comprehension)
- ☐ Guided reading
- ☐ Independent work
- ☐ Small group reading instruction (such as phonics, phonemic awareness, fluency, vocabulary, comprehension)
- ☐ Learning centers
- ☐ Other reading (specify)

Language Arts

- ☐ Spelling
- ☐ Writing (such as writing process, grammar, punctuation, etc.)

Mathematics

- ☐ Whole class math instruction
- ☐ Small group math instruction

- ☐ Physical education
- ☐ Science
- ☐ Social Studies
- ☐ Recess
- ☐ Music
- ☐ Art
- ☐ Computer lab

If the activity/class is not listed, please specify in the space provided below:

Reading

- ☐ Whole class reading instruction such as phonics, phonemic awareness, fluency, vocabulary, comprehension)
- ☐ Guided reading
- ☐ Independent work
- ☐ Small group reading instruction (such as phonics, phonemic awareness, fluency, vocabulary, comprehension)
- ☐ Learning centers
- ☐ Other reading (specify)

Language Arts

- ☐ Spelling
- ☐ Writing (such as writing process, grammar, punctuation, etc.)

Mathematics

- ☐ Whole class math instruction
- ☐ Small group math instruction

- ☐ Physical education
- ☐ Science
- ☐ Social Studies
- ☐ Recess
- ☐ Music
- ☐ Art
- ☐ Computer lab

If the activity/class is not listed, please specify in the space provided below:

Reading

- ☐ Whole class reading instruction such as phonics, phonemic awareness, fluency, vocabulary, comprehension)
- ☐ Guided reading
- ☐ Independent work
- ☐ Small group reading instruction (such as phonics, phonemic awareness, fluency, vocabulary, comprehension)
- ☐ Learning centers
- ☐ Other reading (specify)

Language Arts

- ☐ Spelling
- ☐ Writing (such as writing process, grammar, punctuation, etc.)

Mathematics

- ☐ Whole class math instruction
- ☐ Small group math instruction

- ☐ Physical education
- ☐ Science
- ☐ Social Studies
- ☐ Recess
- ☐ Music
- ☐ Art
- ☐ Computer lab

If the activity/class is not listed, please specify in the space provided below:

Source: Classroom instructional checklist distributed April 2009–May 2009.

Appendix G: Models used for confirmatory, exploratory, and sensitivity analyses

Chapter 2 briefly described the confirmatory, exploratory, and sensitivity analyses models used in this study. This appendix provides the estimation models.

Multiple imputation

Five multiply imputed datasets were created for each group, and then combined to create five overall imputed datasets. Each multiply imputed dataset included the screener composite score, gender, race/ethnicity, free or reduced-price lunch status, Individualized Education Program status, English language learner status, and dummy indicator variables for schools and school pairs to account for the clustered structure of the data. Data analysis was conducted using hierarchical linear modeling (HLM) software version 6.02, which provides results aggregated across the multiply imputed datasets using the rules developed by Rubin (1987).⁸¹

Confirmatory analysis

The primary research question was evaluated using HLM models that compare Test of Early Mathematics Ability–Third Edition (TEMA–3; Ginsburg and Baroody 2003) outcomes of students in the intervention schools with TEMA–3 outcomes of students in the control schools. Specifically, a three-level HLM model was constructed with students at level 1, schools at level 2, and school pairs at level 3. The model is specified as follows:

Level 1 (student level):

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} * (\text{Screen})_{ijk} + e_{ijk}$$

where:

Y_{ijk} is the outcome for student i in school j in pair k .

π_{0jk} is the average outcome of students in school j in pair k when Screen = grand mean.

⁸¹ It should be noted that the imputation process used fixed effects (for schools and pairs), while the treatment effect in the confirmatory and exploratory analyses was estimated with a random effects model. The prediction model used in imputing missing data is essentially a single-level model with schools and pairs as fixed effects (dummies); therefore, some inconsistency exists. However, at the present time, commonly used multiple imputation programs do not accommodate multi-level data. Experts were consulted and were not able to provide alternate solutions. The experts and authors do not believe this issue raises a risk as to the findings in this study.

Screen_{ijk} is the pretest screen score for student i in school j in pair k , grand-mean centered.

π_{ijk} is the relationship of pretest screen to the outcome of student i in school j in pair k .

e_{ijk} is a random error associated with student i in school j in pair k .

$$e_{ijk} \sim N(0, \sigma^2).$$

Level 2 (school level):

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} * (\text{RtI})_{jk} + r_{0jk}$$

$$\pi_{1jk} = \beta_{10k}$$

where:

β_{00k} is the average student outcome across all schools in pair k , adjusted for student pretest screener.

RtI (Response to Intervention) is an indicator variable for the intervention: 1 = RtI ; 0 = Control, group-mean centered.

β_{01k} is the difference in average student outcome between the RTI school and the control school in pair k (that is, intervention effect).

r_{0jk} is a random error associated with school j in pair k on school average student outcome.

$$r_{0jk} \sim N(0, \tau_{00k}).$$

Level 3 (school-pair level):

$$\beta_{00k} = \gamma_{000} + u_{00k}$$

$$\beta_{01k} = \gamma_{010}$$

$$\beta_{10k} = \gamma_{100}$$

where:

γ_{000} is the average student outcome across all pairs (in effect, grand mean).

γ_{010} is the average intervention effect across all pairs.

γ_{100} is the average effect of pretest screener across all pairs.

u_{00k} is a random error associated with pair k on average student outcome, and $u_{00k} \sim N(0, \tau_{000})$.

To determine statistical significance, $\alpha = 0.05$ with a two-tailed test was used. Only one outcome measure (the TEMA–3) and a single confirmatory impact analysis were proposed; therefore, correction for multiple comparisons was not necessary.

In addition to the statistical significance of the RtI effect, the analysis gauges the magnitude of the impact with the effect size index. The following version of Hedges' g , as recommended by the What Works Clearinghouse for calculating effect sizes in cluster randomized studies using HLM analyses, was used (Institute for Education Sciences 2008):

$$g = \frac{\lambda}{\sqrt{\frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{(n_1 + n_2 - 2)}}}$$

where:

λ is the HLM coefficient for the effect due to intervention, which is a group-mean difference adjusted for pretest scores.

In the denominator, n_1 and n_2 are the level 1 sample sizes, and S_1 and S_2 are the unadjusted level 1 standard deviations for the intervention and control group. This formula assumes grand-mean centered variables, the case for the HLM analyses conducted for the present study.

Sensitivity analyses

Six sensitivity analyses were conducted to examine the robustness of the confirmatory impact estimate. These sensitivity analyses use the same HLM model used for the confirmatory impact estimate, with minor adjustments to either the model or the sample, as described below.

The first evaluated whether the impact estimate was sensitive to the exclusion of the 26 schools from the sample that were affected by the natural disaster. Because of the natural disaster, the schedule for those schools differed from those of the other participating schools; 26 schools began *Number Rockets* implementation in January, with four or five lessons delivered per week whenever possible. The HLM model for the confirmatory impact estimate was applied to a sample that included only the other 50 schools that participated in this study.

The second evaluated the decision to explicitly specify the matched school pairs in the HLM model. A two-level HLM analysis was conducted, identical to the first two levels used for the confirmatory impact estimate. (Level 3, which specified school pairs, was not included.)

The third evaluated the robustness of the model to the exclusion of the pretest covariate. This analysis was conducted using the confirmatory impact HLM model, but it did not include the pretest covariate.

The fourth evaluated the sensitivity of the study results to the missing data approach used. The analysis was conducted using the same HLM model used for the confirmatory impact estimate, but it was used on a complete case sample using casewise deletion.

The fifth examined the robustness of the confirmatory impact estimate to the decision to include students not identified as at risk ($n = 45$) in some tutoring groups, to satisfy the study's commitment to provide *Number Rockets* to a minimum of nine students at each intervention school. The analysis was conducted using the confirmatory HLM model and excluded at-risk students ($n = 970$ retained, $n = 24$ excluded) assigned to tutoring groups that had students who were not part of the at-risk analytic sample. In other words, entire student groups that included students not at risk were excluded from the analysis.

The sixth sensitivity analysis also examined this decision by using the confirmatory HLM model and excluding entire school pairs in which any tutoring groups included students who were not part of the at-risk analytic sample. In this analysis, 29 of 38 pairs were retained, resulting in 883 students included in the analysis and 111 students excluded.

See appendix M for the results of all six sensitivity analyses.

Exploratory impact analyses

Exploratory model 1

To address exploratory research question 1, the main impact model was modified to include a cross-level interaction between school treatment status and student pretest screening scores to examine whether the treatment impact differs as a function of baseline mathematics proficiency. The analysis defines the pretest screener score (*Screen*) as a variable representing risk status for each student. (In the study, the pretest screener z -score was used to determine at-risk status.)

The impact of *Number Rockets* for students at various levels of risk, as defined by the pretest screener score, was tested using a three-level HLM model, defined as follows:

Level 1 (student level):

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} * (\text{Screen})_{ijk} + e_{ijk}$$

where:

Y_{ijk} is the outcome for student i in school j in pair k .

π_{0jk} is the average outcome of students in school j in pair k when Screen = grand mean.

Screen_{ijk} is the pretest screener score for student i in school j in pair k , grand-mean centered.

π_{1jk} is the relationship of pretest screener score on the outcome of student i in school j in pair k .

e_{ijk} is a random error associated with student i in school j in pair k and $e_{ijk} \sim N(0, \sigma^2)$.

Level 2 (school level):

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} * (RtI)_{jk} + r_{0jk}$$

$$\pi_{1jk} = \beta_{10k} + \beta_{11k} * (RtI)_{jk} + r_{1jk}$$

where:

β_{00k} is the average student outcome across all schools in pair k .

RtI is an indicator variable for the intervention: 1 = RtI ; 0 = Control, group-mean centered.

β_{01k} is the difference in average student outcome between the RtI school and the control school in pair k (in effect, intervention effect), adjusted for student pretest screener score.

β_{10k} is the average relationship between student pretest screener score and the outcome across all schools in pair k .

β_{11k} is the difference between the RtI school and the control school in the relationship between student pretest screener scores and the outcome in pair k .

r_{0jk} is a random error associated with school j in pair k on school average student outcome.

r_{1jk} is a random error associated with school j in pair k on the relationship between student screen scores and the outcome.

$$\begin{pmatrix} r_{0jk} \\ r_{1jk} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{000} & \tau_{001} \\ \tau_{100} & \tau_{111} \end{pmatrix} \right]$$

Level 3 (school-pair level):

$$\beta_{00k} = \gamma_{000} + u_{00k}$$

$$\beta_{01k} = \gamma_{010}$$

$$\beta_{10k} = \gamma_{100}$$

$$\beta_{11k} = \gamma_{110}$$

where:

γ_{000} is the average student outcome across all pairs (in effect, grand mean).

u_{00k} is a random error associated with pair k on average student outcome and $u_{00k} \sim N(0, \tau_{000})$.

γ_{010} is the average intervention effect across all pairs.

γ_{100} is the average relationship between pretest screener scores and the outcome across all pairs.

γ_{110} is the average intervention effect on the relationship between pretest screener scores and the outcome across all pairs.

Exploratory model 1: sensitivity analysis

The following procedures were conducted to address the research question posed by the sensitivity analysis conducted for exploratory model 1.

First, all at-risk students (both treatment and control, $n = 994$) were ranked by pretest screener scores, and three student-ability groups were created: the lowest third ($n = 331$), the middle third ($n = 331$), and the highest third ($n = 332$). Second, the HLM model used in the primary confirmatory analysis was used to estimate the impact of *Number Rockets* separately for each student-ability groups. Full results are reported in tables M-11 through M-13. Finally, the impact estimate for the lowest third was compared with both the middle and highest third, and the highest third with both the middle and lowest third. Full results are reported in tables M-14 and M-15. The HLM models are as follows.

The sensitivity analysis compared the treatment effect across these three subgroups using the following model:

Level 1 (student level):

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} * (\text{Third2})_{ijk} + \pi_{2jk} * (\text{Third3})_{ijk} + e_{ijk}$$

where:

Y_{ijk} is the outcome for student i in school j in pair k .

π_{0jk} is the outcome of students in school j in pair k when level 1 covariates are set to zero.

Third2 is a dummy variable indicating that student i is in the middle third on the pretest screener.

Third3 is a dummy variable indicating that student i is in the upper third on the pretest screener.

(Third1 is the omitted category for student i being in the lower third on the pretest screener.)

π_{1jk} is the difference between being in the lower versus the middle third of the pretest screener on the outcome of student i when all other covariates in the model are set to zero.

π_{2jk} is the difference between being in the lower versus the upper third of the pretest screener on the outcome of student i when all other covariates in the model are set to zero.

e_{ijk} is a random error associated with student i in school j in pair k and $e_{ijk} \sim N(0, \sigma^2)$.

Level 2 (school level):

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} * (RtI)_{jk} + r_{0jk}$$

$$\pi_{1jk} = \beta_{10k} + \beta_{11k} * (RtI)_{jk}$$

$$\pi_{2jk} = \beta_{20k} + \beta_{21k} * (RtI)_{jk}$$

where:

β_{00k} is the average student outcome across all schools in pair k .

RtI is an indicator variable for the intervention: $1 = RtI$; $0 = \text{Control}$, group-mean centered.

β_{01k} is the difference in average student outcome between the RtI school and the control school in pair k (in effect, overall intervention effect).

β_{10k} is the average relationship between being in the lower versus middle third with student outcome across all classrooms in school k .

β_{11k} is the difference between the *RtI* school and the control school in the relationship between being in the lower versus middle third with the outcome in pair k .

β_{20k} is the average relationship between being in the lower versus upper third with student outcome across all classrooms in school k .

β_{21k} is the difference between the *RtI* school and the control school in the relationship between being in the lower versus upper third with the outcome in pair k .

r_{0jk} is a random error associated with school j in pair k on school average student outcome.

$r_{0jk} \sim N(0, u_{00k})$.

Level 3 (school-pair level):

$$\beta_{00k} = \gamma_{000} + u_{00k}$$

$$\beta_{01k} = \gamma_{010}$$

$$\beta_{10k} = \gamma_{100}$$

$$\beta_{11k} = \gamma_{110}$$

$$\beta_{20k} = \gamma_{200}$$

$$\beta_{21k} = \gamma_{210}$$

where:

γ_{000} is the average student outcome across all pairs (in effect, grand mean).

u_{00k} is a random error associated with pair k on average student outcome and $u_{00k} \sim N(0, \tau_{000})$.

γ_{010} is the average intervention effect across all pairs.

γ_{100} is the average relationship between being in the lower versus middle third with student outcome across all pairs.

γ_{110} is the difference between *RtI* schools and control schools in the relationship between being in the lower versus middle third with the outcome across all pairs.

γ_{200} is the average relationship between being in the lower versus upper third with student outcome across all pairs.

γ_{210} is the difference between *RtI* schools and control schools in the relationship between being in the lower versus upper third with the outcome across all pairs.

Exploratory model 2

To address research exploratory question 2, the main impact model was used, but here the Woodcock-Johnson—Third Edition Letter/Word subtest was used as the outcome. The analysis defines the pretest screener score (*Screen*) as a variable representing risk status for each student. (In the study, the pretest screener z-score is used to determine at-risk status.)

The impact of *Number Rockets* for students at various levels of risk as defined by the pretest screener score was tested using a three-level HLM model, defined as follows:

Level 1 (student level):

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} * (\text{Screen})_{ijk} + e_{ijk}$$

where:

Y_{ijk} is the outcome for student i in school j in pair k .

π_{0jk} is the average outcome of students in school j in pair k when *Screen* = grand mean.

Screen_{ijk} is the pretest screen score for student i in school j in pair k , grand-mean centered.

π_{1jk} is the relationship of pretest screen to the outcome of student i in school j in pair k .

e_{ijk} is a random error associated with student i in school j in pair k .

$$e_{ijk} \sim N(0, \sigma^2).$$

Level 2 (school level):

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} * (\text{RtI})_{jk} + r_{0jk}$$

$$\pi_{1jk} = \beta_{10k}$$

where:

β_{00k} is the average student outcome across all schools in pair k , adjusted for student pretest screener.

RtI is an indicator variable for the intervention: 1 = RtI; 0 = Control, group-mean centered.

β_{01k} is the difference in average student outcome between the RTI school and the control school in pair k (that is, intervention effect).

r_{0jk} is a random error associated with school j in pair k on school-average student outcome.

$$r_{0jk} \sim N(0, \tau_{00k}).$$

Level 3 (school-pair level):

$$\beta_{00k} = \gamma_{000} + u_{00k}$$

$$\beta_{01k} = \gamma_{010}$$

$$\beta_{10k} = \gamma_{100}$$

where:

γ_{000} is the average student outcome across all pairs (in effect, grand mean).

γ_{010} is the average intervention effect across all pairs.

γ_{100} is the average effect of pretest screener across all pairs.

u_{00k} is a random error associated with pair k on average student outcome.

$$u_{00k} \sim N(0, \tau_{000}).$$

Exploratory model 3

To address research exploratory question 3, the main impact model was used but adapted by adding the implementation variable (*Session*) to the school-pair level of the model. Since the control group did not implement the intervention, it was not possible to directly measure implementation level in control schools. However, as the schools were randomly assigned within blocked school pairs, the number of sessions implemented in the treatment school served as the measure of implementation for both schools within each respective pair. Because the implementation level could not vary across schools within a pair, the number of sessions was included as a school-pair characteristic.

The relationship between the level of implementation of *Number Rockets*, as measured by the average number of sessions, and the effect of the intervention on school-pair level impact estimates was tested using a three-level HLM model, defined as follows:

Level 1 (student level):

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} * (Screen)_{ijk} + e_{ijk}$$

where:

Y_{ijk} is the outcome for student i in school j in pair k .

π_{0jk} is the average outcome of students in school j in pair k when $Screen =$ grand mean.

$Screen_{ijk}$ is the pretest screener score for student i in school j in pair k , grand-mean centered.

π_{1jk} is the effect of pretest screener score on the outcome of student i in school j in pair k .

e_{ijk} is a random error associated with student i in school j in pair k and $e_{ijk} \sim N(0, \sigma^2)$.

Level 2 (school level):

$$\pi_{0jk} = \beta_{00k} + \beta_{01k} * (RtI)_{jk} + r_{0jk}$$

$$\pi_{1jk} = \beta_{10k}$$

where:

β_{00k} is the average student outcome across both schools in pair k .

RtI is an indicator variable for the intervention: $+1/2 = RtI$; $-1/2 =$ Control, group-mean centered.

β_{01k} is the difference in average student outcome between the RtI school and the comparison school in pair k (that is, intervention effect), adjusted for student pretest screener score.

r_{0jk} is a random error associated with school j in pair k on school average student outcome and $r_{0jk} \sim N(0, \tau_{00k})$.

β_{10k} is the average relationship between pretest screener scores and the outcome across both schools in pair k .

Level 3 (school-pair level):

$$\beta_{00k} = \gamma_{000} + \gamma_{001} * (Session)_k + u_{00k}$$

$$\beta_{01k} = \gamma_{010} + \gamma_{011} * (Session)_k$$

$$\beta_{10k} = \gamma_{100}$$

where:

γ_{000} is the average student outcome across all pairs (in effect, the grand mean).

γ_{001} is the effect of the average number of *Sessions* delivered within pair k on the average student outcome across both schools in pair k .

$Session_k$ is the average number of *Sessions* delivered within pair k , grand-mean centered.

u_{00k} is a random error associated with pair k on average student outcome and $u_{00k} \sim N(0, \tau_{000})$.

γ_{010} is the average intervention effect across all pairs.

γ_{011} is the estimate of the relationship between the average number of *Sessions* and the pair k treatment effect.

γ_{100} is the average relationship between pretest screener scores and the outcome across all pairs.

When displayed in the combined form, the model is:

$$Y_{ijk} = \gamma_{000} + \gamma_{010} * (RtI)_{jk} + \gamma_{001} * (Session)_k + \gamma_{011} * (Session)_k * (RtI)_{jk} + \gamma_{100} * (Screen)_{ijk} + u_{00k} + r_{0jk} + e_{ijk}$$

Which indicates the inclusion of both the level of implementation main effect, as well as the interaction between the level of implementation and the effect of the intervention at the school-pair level.

Appendix H: Lessons

Table H-1. *Number Rockets*, required and additional lessons by topic and day

<i>Topic</i>	<i>Day</i>	<i>Content</i>	<i>Required or additional</i>	<i>Criteria for determining if additional day's lesson is or is not required^a</i>	<i>If day is skipped proceed to:</i>	<i>Notes</i>
1	1	Identifying and Writing Numbers	Required			
1	2	Identifying and Writing Numbers	Additional	Mastery criteria are met for Day 1: counting concrete objects 9/10 items correct Topic 1, Day 1 tutoring Sheets 2, 3, 4—18/20 items correct writing numbers 18/20 items correct	Topic 2, Day 1	
1	3	Identifying and Writing Numbers	Additional	Mastery criteria are met for Day 2: Topic 1, Day 2 tutoring Sheet 2—10/12 items correct Topic 1, Day 2 tutoring Sheets 3 and 4—18/20 items correct writing numbers to 99: 14/16 items correct	Topic 2, Day 1	Proceed to Topic 2, Day 1 regardless of mastery
2	1	Identifying More and Less Objects	Required			Should assess more, less, equal
2	2	Identifying More and Less Objects	Additional	Mastery criteria are met on Day 1: Topic 2, Day 1 tutoring Sheets 1, 2, 3—4/4 items correct	Topic 3, Day 1	
2	3	Identifying More and Less Objects	Additional	Mastery criteria are met for Day 2: Topic 2, Day 2 tutoring Sheets 3, 4, 5—4/4 items correct	Topic 3, Day 1	Proceed to Topic 3, Day 1 regardless of mastery
3	1	Sequencing Numbers	Required	If students reach mastery on tutoring Sheet 1, move to Day 2, if students haven't reached mastery, move to tutoring Sheet 2 mastery criteria for Topic 3, Day 1 tutoring Sheet 1—8/9 items correct mastery criteria for Topic 3, Day 1 tutoring Sheet 2—15/18 items correct		Note: Topic 3 activities could be completed in 1–2 sessions. If students meet Day 1 mastery, tutor may proceed to Day 2, if mastery is met for Day 2, tutor may proceed to Day 3
3	2	Sequencing Numbers	Required	If students receive 5/6 items correct on Topic 3, Day 2 tutoring Sheet 1, proceed to Day 3		
3	3	Sequencing Numbers	Required			If students do not reach mastery criteria for tutoring Sheet 1 (9/10 items correct), they should complete tutoring Sheet 2
4	1	Using <, >, and =	Required			
4	2	Using <, >, and =	Additional	Mastery criteria are met for Day 1: Topic 4, Day 1 tutoring Sheet 1—6/8 items correct Topic 4, Day 2 tutoring Sheet 2—6/8 items correct	Topic 5, Day 1	
4	3	Using <, >, and =	Additional	Mastery criteria are met for Day 1:	Topic 5,	Proceed to Topic 5, Day 1 regardless of

<i>Topic</i>	<i>Day</i>	<i>Content</i>	<i>Required or additional</i>	<i>Criteria for determining if additional day's lesson is or is not required^a</i>	<i>If day is skipped proceed to:</i>	<i>Notes</i>
				Topic 4, Day 1 tutoring Sheet 1—6/8 items correct Topic 4, Day 2 tutoring Sheet 2—6/8 items correct	Day 1	mastery
5	1	Skip Counting by 10s, 5s, and 2s	Required			
5	2	Skip Counting by 10s, 5s, and 2s	Additional	Mastery criteria are met for Day 1: Topic 5, Day 1 tutoring Sheet 2—7/8 items correct Topic 5, Day 1 tutoring Sheet 4—7/8 items correct Topic 5, Day 1 tutoring Sheet 6—7/8 items correct	Topic 6, Day 1	
5	3	Skip Counting by 10s, 5s, and 2s	Additional	Mastery criteria are met for Day 1: Topic 5, Day 1 tutoring Sheet 2—7/8 items correct Topic 5, Day 1 tutoring Sheet 4—7/8 items correct Topic 5, Day 1 tutoring Sheet 6—7/8 items correct or Mastery criteria are met for Day 2: Topic 5, Day 1 tutoring Sheet 2—7/8 items correct Topic 5, Day 1 tutoring Sheet 4—7/8 items correct Topic 5, Day 1 tutoring Sheet 6—7/8 items correct	Topic 6, Day 1	
6	1	Introduction to Place Value	Required			Mastery criteria: Topic 6, Day 1 tutoring Sheet 1— 15/18 items correct
6	2	Introduction to Place Value	Required			Mastery criteria: Topic 6, Day 2 tutoring Sheet 2— 15/18 items correct
6	3	Introduction to Place Value	Required			Mastery criteria: Topic 6, Day 3 tutoring Sheets 2, 3, 4, 5—12/15 items correct
7	1	Place Value	Required			If the student reaches mastery criteria for Day 1 and the tutor feels the student could be quicker, tutor may continue with Day 3
7	2	Place Value	See notes	Mastery criteria are met for Day 1: Topic 7, Day 1 tutoring Sheet 2—9/9 items correct		Day 2 objectives must be covered, based on tutor discretion Day 2 objectives may be covered in Day 1 mastery criteria: Topic 7, Day 2 tutoring Sheet 2—

<i>Topic</i>	<i>Day</i>	<i>Content</i>	<i>Required or additional</i>	<i>Criteria for determining if additional day's lesson is or is not required^a</i>	<i>If day is skipped proceed to:</i>	<i>Notes</i>
						9/9 items correct
7	3	Place Value	Required			Day 3 objectives must be covered, based on tutor discretion Day 3 objectives may be covered in Day 1 or Day 2 Mastery criteria: Topic 7, Day 3 tutoring Sheet 3—9/9 items correct
8	1	Identifying Operations	Required			
8	2	Identifying Operations	Additional	Mastery criteria are met for Day 1: Topic 8, Day 1 tutoring Sheet 2—16/18 items correct	Topic 9, Day 1	
8	3	Identifying Operations	Additional	Mastery criteria are met for Day 1: Topic 8, Day 1 tutoring Sheet 2—16/18 items correct or mastery criteria are met on Day 2: Topic 8, Day 2 tutoring Sheet 1—16/18 items correct or Topic 8, Day 2 tutoring Sheet 2	Topic 9, Day 1	Proceed to Topic 9, Day 1 regardless of mastery after 3 days with Topic 8 mastery criteria: Topic 8, Day 3 tutoring Sheet 2—16/18 items correct Topic 8, Day 3 tutoring Sheet 2—16/18 items correct
9	1	Writing Addition and Subtraction Sentences	Required			Proceed to Day 2 activities if mastery is reached for Day 1 tutoring Sheet 2 mastery criteria: Topic 9, Day 1 tutoring Sheets 2 through 9—5/6 items correct
9	2	Writing Addition and Subtraction Sentences	Required			All students must complete Day 2 regardless of mastery of Day 1 mastery criteria: Topic 9, Day 2 tutoring Sheet 2—5/6 items correct
9	3	Writing Addition and Subtraction Sentences	Additional	Mastery criteria are met for Day 2: Topic 9, Day 2 tutoring Sheet 2—16/18 items correct	Topic 10, Day 1	Mastery criteria: Topic 9, Day 3 tutoring Sheet 3—5/6 items correct
10	1	Place Value	Required			Mastery criteria: Topic 10, Day 1 tutoring Sheet 2—8/10 items correct

<i>Topic</i>	<i>Day</i>	<i>Content</i>	<i>Required or additional</i>	<i>Criteria for determining if additional day's lesson is or is not required^a</i>	<i>If day is skipped proceed to:</i>	<i>Notes</i>
10	2	Place Value	Required			Mastery criteria: Topic 10, Day 2 tutoring Sheet 2—8/10 items correct
10	3	Place Value	Required			Mastery criteria: Topic 10, Day 3 tutoring Sheet 2—12/15 items correct
11	1	Addition Facts	Required			Note: Students must spend a minimum of 4 days on Topic 11 mastery criteria: Topic 11, Day 1 tutoring Sheets 1–10, 100 percent mastery for all 10 sheets
11	2	Addition Facts	Additional	Mastery criteria are met for Day 1: Topic 11, Day 1 tutoring Sheets 1–10, 100 percent mastery for all 10 sheets	Topic 11, Day 4	
11	3	Addition Facts	Additional	Mastery criteria are met for Day 1: Topic 11, Day 1 tutoring Sheets 1–10, 100 percent mastery for all 10 sheets	Topic 11, Day 4	
11	4	Addition Facts	Required			No mastery criteria, tutor selects review activity for the day
11	5	Addition Facts	Required			No mastery criteria, tutor selects review activity for the day
11	6	Addition Facts	Required			No mastery criteria, tutor selects review activity for the day
12	1	Subtraction Facts	Required			Note: Students must spend a minimum of 4 days on Topic 11 mastery criteria: Topic 12, Day 1 tutoring Sheets 1–10, 100 percent mastery for all 10 sheets
12	2	Subtraction Facts	Additional	Mastery criteria are met for Day 1: Topic 12, Day 1 tutoring Sheets 1–10, 100 percent mastery for all 10 sheets	Topic 12, Day 4	
12	3	Subtraction Facts	Additional	Mastery criteria are met for Day 1: Topic 12, Day 1 tutoring Sheets 1–10, 100 percent mastery for all 10 sheets	Topic 12, Day 4	
12	4	Subtraction	Required			No mastery criteria, tutor selects

<i>Topic</i>	<i>Day</i>	<i>Content</i>	<i>Required or additional</i>	<i>Criteria for determining if additional day's lesson is or is not required^a</i>	<i>If day is skipped proceed to:</i>	<i>Notes</i>
		Facts				review activity for the day
12	5	Subtraction Facts	Required			No mastery criteria, tutor selects review activity for the day
12	6	Subtraction Facts	Required			No mastery criteria, tutor selects review activity for the day
13	1	Addition and Subtraction Facts Review	Required			Mastery criteria: Topic 13, Day 1 tutoring Sheets 1, 2, 3—80 percent or 44/54 items correct on at least one of the sheets
13	2	Addition and Subtraction Facts Review	Required			Mastery criteria: Topic 13, Day 2 tutoring Sheets 1, 2, 3—80 percent or 44/54 items correct on at least one of the sheets
13	3	Addition and Subtraction Facts Review	Required			Mastery criteria: Topic 13, Day 3 tutoring Sheets 1 and 2—80 percent or 44/54 items correct on at least one of the sheets
14	1	Place Value Review	Required			Mastery criteria: Topic 14, Day 1 tutoring Sheets 2, 3, or 4—10/12 items correct
14	2	Place Value Review	Required			Mastery criteria: Topic 14, Day 2 tutoring Sheets 2, 3, or 4—10/12 items correct
14	3	Place Value Review	Required			Mastery criteria: Topic 14, Day 3 tutoring Sheets 2, 3, or 4—10/12 items correct
15	1	2-Digit Addition	Required			Note: Students must spend a minimum of 3 days in this topic No mastery criteria
15	2	2-Digit Addition	Required			No mastery criteria
15	3	2-Digit Addition	Required			
15	4	2-Digit Addition	Additional	Mastery criteria are met for Day 3: Topic 15, Day 3 tutoring Sheet 1 or 2—14/16 items	Topic 16, Day 1	

<i>Topic</i>	<i>Day</i>	<i>Content</i>	<i>Required or additional</i>	<i>Criteria for determining if additional day's lesson is or is not required^a</i>	<i>If day is skipped proceed to:</i>	<i>Notes</i>
				correct and students are able to complete the tutoring sheets without assistance or Base-10 blocks		
15	5	2-Digit Addition	Additional	Mastery criteria are met for Day 4: Topic 15, Day 4 tutoring Sheet 1 or 2—14/16 items correct	Topic 16, Day 1	
15	6	2-Digit Addition	Additional	Mastery criteria are met for Day 5: Topic 15, Day 5 tutoring Sheet 1 or 2—14/16 items correct	Topic 16, Day 1	Note: move to Topic 16 even if mastery has not been met
16	1	2-Digit Subtraction	Required			Note: Students must complete a minimum of 3 days in this topic No mastery criteria for Day 1
16	2	2-Digit Subtraction	Required			No mastery criteria
16	3	2-Digit Subtraction	Required			
16	4	2-Digit Subtraction	Additional	Mastery criteria are met for Day 3: Topic 16, Day 3 tutoring Sheet 1 or 2—14/16 items correct and students are able to complete tutoring sheets without assistance or Base-10 blocks	Topic 17, Day 1	
16	5	2-Digit Subtraction	Additional	Mastery criteria are met for day 4: Topic 16, Day 4 tutoring Sheet 1 or 2—14/16 items correct	Topic 17, Day 1	
16	6	2-Digit Subtraction	Additional	Mastery criteria are met for day 5: Topic 16, Day 5 tutoring Sheet 1 or 2—14/16 items correct	Topic 17, Day 1	Mastery criteria: Topic 16, Day 6 tutoring Sheet 1 or 2—14/16 items correct
17	1	Missing Addends	Required			Mastery criteria: Topic 17, Day 1 tutoring Sheet 2—16/20 items correct
17	2	Missing Addends	Required			Mastery criteria: Topic 17, Day 2 tutoring Sheet 2—16/20 items correct
17	3	Missing Addends	Required			Mastery criteria: Topic 17, Day 3 tutoring sheet 2—16/20 items correct

a. The *Number Rockets* intervention covers 17 topics, each divided into three to six lessons. Not all lessons are required. If the entire group of students meets the mastery criteria for a topic during a required lesson, the additional days/lessons for the topic are skipped. Although the entire intervention can be completed in as few as 41 lessons, students still cover all 17 topics regardless of the number of lessons skipped due to meeting mastery criteria.

Source: Paulsen and Fuchs 2005; authors' summary of *Number Rockets* intervention.

Appendix I: Complete sample lesson Topic 6, Day 1

The first activity is a review of material covered in the previous Topic 5—*Skip Counting by 10s, 5s, and 2s*. The tutor begins reading from the beginning of the script (figure I-1).

Figure I-1. Excerpt from *Number Rockets* tutoring script





Tutor: The first thing we need to do today is complete this review sheet. I'll read the questions and you write the answers.

Action: Read directions and allow time for students to answer.

Source: Paulsen and Fuchs (2005, p. 57).

Next, students complete a review sheet (figure I-2).

Figure I-2. Review sheet #5

Review #5		Name: _____
<p>①</p> <p>Counting by 5's. Fill in the blanks:</p> <p>15, 20, 25, _____, _____</p>	<p>⑤</p> <p>What number comes after 26?</p> <p>26 _____</p>	
<p>②</p> <p>Counting by 10's. Fill in the blanks:</p> <p>20, 30, 40, _____, _____</p>	<p>⑥</p> <p>Circle the set that has more than .</p> <p>  </p>	
<p>③</p> <p>Fill in the blanks:</p> <p>75, 76, 77, _____, _____</p>	<p>⑦</p> <p>Counting by 5's. Fill in the blanks:</p> <p>50, 55, 60, _____, _____</p>	
<p>④</p> <p>Counting by 10's. Fill in the blanks:</p> <p>50, 60, 70, _____, _____</p>	<p>⑧</p> <p>Write <, >, or = in the blank.</p> <p>24 _____ 47</p>	

Source: Paulsen and Fuchs 2005.

To begin the Topic 6, Day 1 lesson, the tutor presents the concept of *place value* verbally and writes an example (figure I-3). Note how positive feedback to students such as “great work” and “that’s right” is also scripted, as well as corrective feedback such as “these numbers are different from each other because . . .” After the example, each student is provided a worksheet to write answers for the rest of the lesson.

Figure I-3. Excerpt from Topic 6, Day 1 lesson *Number Rockets* script: tutor introduces place value

Great work. Today we're going to be working on place value.

Write the numbers 5 and 13.

How are these numbers different?

If the student gives an incorrect response say, These numbers are different from each other because the 13 takes up two places, but the 5 only takes up one place. How are the numbers different?

Students should respond something like:

5 takes up one place

13 takes up two places

That's right. These numbers are different from each other because the 13 takes up two places; two numbers together make up 13.

But 5 only takes up one place. So, 5 takes up one place, but 13 takes up two places.

Give students Topic 6 Day 1 Tutoring Sheet 1.

These places have a special name. Write 5 in the ones place of the first box on Topic 6 Day 1 Tutoring Sheet 1. This (point to the 5) is called the ones place. Five only has one place. Write 5 on your sheets. Show me 5 with your fingers.

Students should show five fingers.

Source: Paulsen and Fuchs (2005, p. 57).

After the tutor introduces a verbal example of place value, the tutor demonstrates one way to represent place value with fingers (figure I-4).

Figure I-4. Excerpt from Topic 6, Day 1 lesson *Number Rockets* script: tutor demonstrates place value

Great. Write 13 in the second box. Look at 13. In 13, the 3 is in the ones place. (point to the 3 in 13). Every number has something in the ones place. But, look, 13 takes up two places. (point to the 1 in 13). This is called the tens place.

Now I'm going to show you how to show 13 with your fingers. When we have a number that's in both the ones and tens place we'll "flash" all 10 fingers and then count the ones. Let me show you what I mean.

Flash 10 and count up 11,12,13.

Now you show me 13 with your fingers.

Great work.

Source: Paulsen and Fuchs (2005, p.58).

Next, the tutor introduces Base-10 blocks as another way to teach place value concepts and links it back to both the written and finger representations (figure I-5).

Figure I-5. Excerpt from Topic 6, Day 1 lesson *Number Rockets* script: tutor introduces Base-10 blocks

Give each student a set of Base-10 Blocks.

These are called Base-10 Blocks. You can see that you have cubes (show a cube) and rods (show a rod).

Let's see how we can use these to help us in math.

Point to cubes. These are called cubes. Each cube stands for 1.

Put 8 cubes in front of you and count them. I have 1,2,...8 cubes here. Let's write 8 in the ones place. Allow time to write 8. How many places are in 8?

Students should say one.

Right, the number 8 only has one place. Show me 8 with your fingers.

Students should show 8 fingers.

Source: Paulsen and Fuchs (2005, p. 58).

Additional guided practice⁸² is given, structured as described above, relating unit-blocks to the 10-unit rod and relating blocks to both finger and written representations of place value on the worksheet. The lesson continues with repeated practice of translating 14 more numbers in visual representations using the Base-10 blocks (figure I-6).

Figure I-6. Excerpt from Topic 6, Day 1 lesson: tutor represents numbers, points awarded

Continue this process with 12, 9, 18, 2, 11, 4, 19, 1, 13, 5, 14, 3, 16, and 17.

You've all worked hard today (or other feedback that may be needed), **it's now time to fill in your point sheets.**

3 Points: 15-18 correct answers

2 Points: 11-14 correct answers

1 Point: 7-10 correct answers

If there is extra time, practice counting rods and cubes

Proceed to Flashcard Activity for final 10 minutes of session.

Note: To assist with behavior management, students received award points for mastering lesson content and for positive behavior, meaning all members of the group were on task (defined as “listening carefully, working hard, and following directions.”) When a student earned points they were allowed to choose a small reward (such as a small toy car, keychain, or pencil eraser).

Source: Paulsen and Fuchs (2005, p. 60).

If the next lesson was not required, students would complete a mastery worksheet to determine if they could skip additional lesson(s) for the topic. If the mastery criterion was met by all members of the group, the additional lesson(s) would be skipped.

The final 10 minutes of each lesson consist of mathematics fact practice with varying levels of flashcards.⁸³ The tutor works with one student at a time, while the other students watch. The tutor administers the flashcards to the selected student for a one-minute timed period, and the student responds to as many flashcards as he or she can within that period, taking as much time as needed for each card. If a student responds to one of the cards incorrectly, the tutor leads him or her through a hand-counting procedure to answer the problem. Once the minute is up for that student, the tutor continues around the circle, taking turns with each student. A second round is conducted, during which each student attempts to correctly answer more cards than he or she did in the first round. If the lesson takes longer than planned, the flashcard activity is truncated to keep the total session time within approximately 40 minutes.

⁸² *Guided practice* refers to skill practice facilitated by the tutor.

⁸³ The tutor is prepared with one deck of flashcards for each child, depending on that child's current skill level with addition and subtraction facts.

Figure I-7 provides an example of the student worksheet that would have been completed by each individual student in the course of the Topic 6, Day 1 lesson.

Figure I-7. Topic 6, Day 1 lesson tutoring sheet 1

Tens		Ones		Tens		Ones		Topic 6 Day 1 Tutoring Sheet 1	

Source: Paulsen and Fuchs 2005.

Appendix J: Details of tutor training

To enable tutors to implement *Number Rockets* with fidelity, the training was multifaceted and included an initial one-day training and two two-hour follow-up trainings. The initial training consisted of six sections:

1. Overview of the study and training.
2. Discussion of the structure of *Number Rockets*.
3. Demonstration of *Number Rockets* sequence and techniques.
4. Guided practice through video.
5. Tutor practice.
6. Debriefing and logistics.

Both of the follow-up trainings consisted of instructional tips and frequently asked questions.

Initial training

1) Overview of the study and training

The initial one-day training began with an overview of the *Number Rockets* program, including an explanation of the study purpose, the research questions, the study's at-risk student sample, and the nature of the intervention and control conditions. Tutors also learned some basic information about the program.

2) Discussion of the structure of Number Rockets

This section exposed the tutors to the program's structure, unique elements, and instructional materials. Discussion included the number of lessons, topics, their lengths, and other details of implementation. Tutors were introduced to the 17 topics in the program, which address number concepts, numeration, computation, and story problems. The program elements presented included the lesson structure, the importance of fidelity to the implementation elements, lesson scripting, awarding of points and rewards, use of flashcards, mastery criteria, and data recording sheets. The materials portion of the training was designed to familiarize the tutors with the teacher's manual, the supporting materials, and the manipulatives (for example, Base-10 blocks, ones blocks).

3) Demonstration of Number Rockets sequence and techniques

This section of the training taught tutors how to prepare and use the essential aspects of the lessons, and proper implementation of several lessons was demonstrated. Using Topic 1, Day 1: Identifying and Writing Numbers, the district coach played the

role of the tutor, stepping in and out of the role to explain some of the essential aspects of the lesson while the observing district coach played the role of the student. During training, tutors were also given time to read through the lesson and record questions or concerns. At the end of the demonstration, the district coach addressed tutors' questions and concerns.

The demonstration covered the mastery criteria, planning/delivery checklist, importance of using the prescribed intervention script and corrections, and behavioral expectations and use of rewards.

a) Mastery criteria

To demonstrate mastery, students were expected to correctly answer a certain number of problems. This determined whether the tutor could move to a new topic, skipping additional lessons on the topic that had not been covered or, if needed, remain in the lesson sequence as prescribed. (Mastery criteria are listed at the beginning of each lesson.) Tutors were instructed that mastery criteria must be achieved for all students in the group for any additional lessons to be skipped. For example, tutors learned that if there were three students in the group, and two of them met the mastery criteria on all activities but the third student did not, then the group must continue along the lesson sequence within that same topic.

b) Planning/delivery checklist

Tutors were also given a set of instructions for planning and conducting their lessons. They were directed to follow a sequence of 10 steps:

1. Begin with a quick reminder about behavior expectations.
2. If beginning a new topic, conduct the review exercise for the previous topic.
3. Conduct the lesson giving students guided practice with the skill, following the script precisely.
4. Give students the independent practice sheet and practice the first problem together.
5. Award points for on-task behavior and for answering mathematics problems correctly.
6. Conduct fact practice.
7. Dismiss students.
8. Check the independent practice tutoring sheet against the mastery criteria.
9. Use the results of the mastery criteria to determine next lesson.
10. Prepare next day's lesson.

c) Importance of using prescribed intervention script and corrections

As the training progressed, the district coach also emphasized that the lessons are explicitly scripted so that there is consistency in the instruction provided to students. District coaches also recommended that tutors prepare for their lessons by carefully reading the scripts and highlighting sections of the lesson they would want to emphasize or marking places in the script that would help them maintain a good pace for the lesson. Substantial time was also spent teaching tutors about the correction procedure recommended by the program. The correction procedure is as follows: *Student gives wrong answer, tutor gives correct answer, the tutor repeats the question, and asks student to respond with the correct answer, and the tutor restates the correct answer then asks the student to repeat the answer. Then the tutor gives positive reinforcement (for example, “Great!”, “Terrific!”, “Super!”) followed by restating the correct answer.*

d) Behavioral expectations and use of rewards

District coaches gave helpful hints to tutors about setting behavioral expectations and using the program’s reward system. For example, coaches indicated that having all individual student materials prepared and readily available in advance of the lesson would reduce behavior problems and increase learning time. Regarding rewards, students earned points when they exhibited appropriate behaviors and did well on their independent work. When a student earned a point, it was indicated on the Math Tutoring Point Sheet. When the students received all the possible points on their Math Tutoring Point Sheet, they earned a prize. The tutors award all students in the group a point if all were listening and on task. If one student was off task, no students were awarded points. If a student(s) was off task, the tutor would describe the off-task behavior. The tutor would then explain why no one got a point and suggest that they work harder to stay on task. Students could also earn individual points based on their performance on the independent practice student worksheet. They could earn up to 3 points every day for correct answers on their independent practice student instruction worksheets. Every lesson required students to complete at least one independent practice student worksheet.

4) Guided practice through video

After modeling a lesson, the coaches moved into the guided practice phase of instruction in which tutors practiced teaching the lessons. For this portion of the training, the coaches began by showing tutors a DVD of Topic 2, Day 1 lesson: Identifying More and Less Objects. Tutors were given a fidelity protocol listing the key behaviors the tutor should implement when teaching the lesson. At predetermined intervals, the trainers stopped the DVD to ask the tutors if they thought the behavior was evident. The coaches highlighted specific tutor behaviors that were the desired ones and those that needed improvement.

Next, the tutors were shown a video on the components of the flashcard activity. Tutors learned the two ways the program teaches students to answer mathematics facts: one was to know facts by memory; the other was to count using the open and closed hand strategies for addition and subtraction. Tutors practiced the strategies with a partner. Time was also provided for tutors to learn about organizing a set of flashcards for each student and award points for flashcard proficiency.

5) Tutor practice

The initial training ended with tutors practicing in trios with one tutor acting as the teacher, a second acting as the student, and a third using the fidelity form to assess for lesson fidelity. The trio practice continued until all tutors practiced each role. During this time, tutors learned and practiced the nuances of implementing several different lessons.

6) Debriefing and logistics

A debriefing session was held at the end of the initial training. Tutors were instructed to resolve any scheduling or space conflicts directly with classroom teachers and school personnel. Tutoring sessions missed were to be rescheduled for the next possible time. On rare occasions where issues could not be resolved by the tutor, they were told to refer matters to the study team, who would communicate directly with the school principal or district contact. Tutors were also given information about how to access support, when the follow-up trainings would be held, and their expectations and responsibilities.

Follow-up trainings

The two follow-up trainings each consisted of two components: 1) instructional tips and 2) frequently asked questions. The instructional tips included recommendations for improving lesson delivery and were developed by the district coaches who listened to audiotaped lessons of each tutor. The coaches listened for patterns and trends among the tutors in instructional strengths and weaknesses and customized their follow-up training agendas on the basis of this analysis. The frequently asked questions focused on three areas: flashcards, lesson pacing, and behavior management.

Appendix K: Tutor background survey

MATH RTI TUTOR BACKGROUND SURVEY

The Department of Education wants us to document and report basic demographics and the experience level of the tutors providing the intervention. This will help schools and districts in the future determine what type of individuals they may wish to recruit to undertake an intervention such as this. All data will be aggregated and **reported at the district-level only!**

No one will be identified individually in any way. We ask for a name only to know from whom we still require a response.

Name: _____ District: _____

Gender: **M** **F**

Ethnicity:

_____ African American _____ Hispanic _____ Asian American/Pacific
Islander

_____ White/Caucasian _____ Native American

Age: 21–24 _____ 25–34 _____ 35–44 _____ 45–54 _____ 55+ _____

Total years teaching: _____

Total years teaching: General education: _____ Special education: _____

Do you typically work as a substitute teacher? **Y...N** (circle one)

Are you currently retired from full-time teaching or full-time substitute teaching? **Y...N** (circle one)

Description of most recent position or current position (e.g., substitute teacher)

Years in Current Position: _____

Education degrees (select all that apply): _____ B.S./B.A./B.Ed. _____ M.S./M.A./M.Ed.
 _____ Ed.S. _____ Ed.D./Ph.D.

Please list all degrees in any other area(s) {e.g. B.A. Psychology}:

Type of teaching certificate held:

_____ Regular or standard

_____ Other (PLEASE SPECIFY) _____

_____ None

Content area of teaching certificate:

_____ Elementary education

_____ Early childhood or

_____ K–12 education

Grade level for teaching certificate: _____ Elementary grades _____ Elementary and secondary grades

Areas of specialization (select all that apply):

_____ Elementary education _____ Early childhood education _____ Special education

_____ Reading _____ Math _____ Other (please specify):

Additional experience: Please describe any other education or child-related experience that you had prior to serving as a tutor in this study (e.g., Reading First instructor, day-care provider, parent)

Source: Partially adapted from Agodini et al. (2009).

Table K-1. Characteristics of mathematics tutors who completed the tutor background survey ($n = 75$), across all districts

<i>Tutor characteristics</i>	<i>n^a</i>	<i>Percentage</i>
Age		
21–34	27	36.0
35–44	— ^e	— ^e
45–54	13	17.3
55+	24	32.0
Not provided	— ^e	— ^e
Gender		
Male	— ^e	— ^e
Female	64	85.3
Not provided	— ^e	— ^e
Race/ethnicity		
White	35	46.7
Black	32	42.7
Hispanic/ Asian/American Indian/Other	8	10.7
Education		
Highest degree earned		
Bachelor's degree	52	69.3
Master's degree or higher	23	30.7
Field for bachelor's degree ^b		
Education	43	57.3
Mathematics/Sciences	6	8.0
Social Sciences	5	6.7
Liberal and Fine Arts	7	9.3
Business/Public Policy	10	13.4
Interdisciplinary studies, general studies/Not provided	4	5.4
Teaching experience (years)		
0–5	33	44.0
6–10	11	14.7
11–15	6	8.0
16–25	5	6.6

<i>Tutor characteristics</i>	<i>n^a</i>	<i>Percentage</i>
26–30	4	5.3
31+	12	16.0
Not provided	4	5.3
Teaching status		
Retired teachers (not substituting)	19	25.3
Substitute teachers (not retired)	29	38.7
Tutors reporting both retired and substitute teacher status	10	13.3
Not retired and not a substitute teacher	17	22.7
Type of teaching certificate held		
Regular or standard	35	46.7
Other	10	13.3
Regular or standard and other	6	8.0
None ^c	15	20.0
Not provided	9	12.0
Content area of teaching certificate		
Elementary education	20	26.7
Early childhood <i>or</i> K-12 education	15	20.0
Elementary education <i>and</i> early childhood	10	13.3
Elementary education and/or early childhood; <i>and</i> K-12 education	8	10.7
Not provided	22	29.3
Area of specialization ^d (tutor could select more than one item)		
Elementary education	32	42.7
Early childhood	15	20.0
Special education	11	14.7
Mathematics/Science	8	10.7
Reading	10	13.3
Language arts	7	9.3

<i>Tutor characteristics</i>	<i>n^a</i>	<i>Percentage</i>
Social studies/Music	9	12.0
Not provided	21	28.0

Note: $n = 75$; within category totals may not sum to 100 because of rounding.

- a. The final tutor sample size was 86. Eleven tutors did not complete the background survey. Three of the 86 served as alternate tutors and did not actively participate in the implementation of the intervention. An alternate tutor is an individual hired and trained to substitute for a current tutor in the event the current tutor was not able to complete intervention implementation.
- b. Education includes one of the following major fields of study: education, education library science, or music education; social science includes one of the following major fields of study: counseling, human services/social work, or psychology; liberal and fine arts includes one of the following major fields of study: art, communication, English, philosophy, history, or Spanish; business includes one of the following major fields of study: accounting, finance, or general business; sciences includes one of the following major fields of study: biology, information technology, engineering, or chiropractic; public policy includes one of the following major fields of study: criminal justice or park administration.
- c. The 15 tutors in this category met the minimum state qualifications and certifications to work as a paraprofessional in the respective state's school system.
- d. Twenty-seven tutors selected one area of specialization; 18 selected two areas of specialization; 6 selected three areas of specialization; and 3 selected four areas of specialization.
- e. Cell counts masked because one or more values within each category are less than three, thereby representing a disclosure risk.

Source: Tutor background surveys completed April 2009–May 2009.

Appendix L: Details of fidelity coder training

Coder training consisted of four components: orientation to the *Number Rockets* structure, elements, and materials; information about the flashcard procedure; practice coding sample lessons; and logistics of completing lesson fidelity checklists for the assigned lessons.

Component 1: orientation to the program structure, elements, and materials

Coders were informed that the program consisted of 17 topics addressing number concepts, numeration, computation, and story problems. Next, they were given a brief overview of the program elements, which included the lesson structure, the importance of fidelity to the program, lesson scripting, point awarding, flashcards, mastery criteria, and data recording sheets. The materials portion of the training provided an overview of the tutor manual, the supporting materials (such as the independent practice student worksheets, review sheets, behavior forms, and flashcards) and the manipulatives (for example, Base-10 blocks and ones blocks). Coders were told that tutors were asked to read from scripts in the tutor manual to ensure consistency in the instruction provided to students. Coders were also told that tutors were instructed to follow specific steps when teaching a lesson.

Coders were informed they would be evaluating the tutor's fidelity to the program and not the tone of their interaction with students. Coders were also advised that tutors were expected to implement a behavior management system in which students earned points when they exhibited appropriate behaviors or did well on their independent practice student worksheets. (Tutors were to record points on a Math Instruction Point Sheet). Coders also learned that tutors would use timers as part of the behavior management system and were expected to determine if all students were on task with three criteria when the timer went off: listening carefully, working hard, and following directions. Coders were told that tutors should have reminded students about the rules for the session and criteria for earning points. The reminders were to be brief, to the point, and not to take more than a minute. Coders learned that most lessons call for the tutor to assign an independent practice student worksheet, and coders were trained to attend to this behavior. Coders were assigned to senior staff who were available to answer questions.

Component 2: information about the flashcard procedure

Coders were presented with a segment on the procedure for the flashcard activity. They learned the two ways tutors would be teaching students mathematics facts. One was to know the addition and subtraction facts by memory. The other was to count using the open- and closed-hand strategies. It was important that coders listened to determine if tutors prompted students to use these strategies during counting practice.

Component 3: practice coding sample lessons

The most important portion of the training for coders was learning and practicing the lesson fidelity checklists (see figure F-1). These lesson fidelity checklists provided evidence on whether the tutors were implementing the essential aspects of lessons. To that end, the training provided guided practice in which coders viewed a videotaped lesson and practiced coding using the lesson fidelity checklists.

Specifically, the trainers began by showing a video of the Topic 2, Day 1 lesson: Identifying More and Less Objects. Coders were given a lesson fidelity checklist identifying the key behaviors a tutor should implement when teaching. The coders were asked to determine whether they observed the behavior. At frequent intervals the trainers would stop to ask the coders if they thought there was evidence of the specific behaviors. The trainers also highlighted specific tutor behaviors that met the criteria for adequate implementation.

The key features of these lessons were highlighted by stopping the videotape every 10 minutes to discuss key teaching behaviors that were correctly performed, incorrectly performed, or missing. Examples included a discussion about the presentation of review sheets, the correction procedure, tutors' demonstrations/modeling of a strategy, monitoring students' work on the independent practice student worksheets, and evidence of use of behavior modification strategies. Coders were cautioned that some items on the checklist required two teaching behaviors. In these cases, the tutor had to implement both behaviors to receive a positive rating.

For most independent practice student worksheets, the tutor manual instructs the tutor to distribute the worksheet, explain the directions, complete one problem as a group, and ask students to complete the remaining problems on their own. Coders were told that in some instances the directions indicate that the tutor should continue to practice the concept by working through all of the problems on the independent practice student worksheet and asking questions to guide students in correctly answering the problems. In these instances, tutors should not have students work independently. The trainers stressed that tutors should have followed the instructions in the tutor manual. Coders then coded three additional practice lessons with the accompanying flashcard activity. After each lesson, the trainer conducted a debriefing.

Component 4: logistics of completing lesson fidelity checklists of the assigned lessons

If an audio recording of a lesson was unavailable, coders were told they would be assigned an alternate lesson and given the accompanying lesson fidelity checklist. Coders were also given directions on how to enter the identifying information at the top of each checklist. This included the coder's name, coding date, tutor's name, school district, and lesson identifier. Coders were also directed to indicate if the lesson was less than 30 minutes by placing a check next to the phrase, "Less than 30 minutes." Coders were told to rate each item on the following scale:

- A checkmark in the + column if the behavior is present.
- A checkmark in the – column if the behavior is not present.
- A checkmark in the N/A column if the item is not applicable.

Coders were also asked to calculate fidelity by dividing the number of checkmarks in the + column by the number of checkmarks in the + and – columns and to enter the score in the designated area at the bottom of the fidelity protocol. They were asked to report the percentage to the tenths place.

For the flashcard activity, coders were instructed to put an asterisk where the activity started and ended. Fidelity was calculated by dividing the number of checkmarks in the + column by the number of checkmarks in the + and – columns for the items between the two asterisks. Coders were directed to enter the score in the designated area at the bottom of the fidelity protocol and report the percentage to the tenths place.

Appendix M: Complete multilevel model results for chapter 4 (confirmatory and sensitivity) and chapter 5 (exploratory and sensitivity) analyses

Analyses reported in chapter 4

Table M-1. Confirmatory impact analysis

<i>Fixed effects model</i>	<i>Coefficient</i>	<i>Standard error</i>	<i>t</i>	<i>Degrees of freedom</i>	<i>p-value</i>
<i>Intercept, γ_{000}</i>	86.18	(0.48)	178.12	37	< .001
<i>Treatment, γ_{010}</i>	4.28	(0.82)	5.24	55	< .001
<i>Screener, γ_{100}</i>	13.63	(0.89)	15.38	269	< .001
<i>Random effects</i>	<i>Standard deviation</i>	<i>Variance component</i>	<i>Degrees of freedom</i>	χ^2	<i>p-value</i>
<i>Student, e_{ijk}</i>	(9.82)	96.44			
<i>School, σ_{jk}</i>	(0.75)	0.57	37	46.99	.126
<i>Pair, u_{00k}</i>	(1.69)	2.85	37	64.00	.004

Note: A three-level hierarchical linear model and five multiply imputed datasets were used to estimate the statistics in this table. Student sample sizes: $n_{total} = 994$; $n_{intervention} = 615$, $n_{control} = 379$. School sample sizes: $n_{total} = 76$; $n_{intervention} = 38$, $n_{control} = 38$.

Source: Study data collected August 2008–May 2009.

Table M-2. Sensitivity analysis 1: excluding 26 schools affected by natural disaster

<i>Fixed effects model</i>	<i>Coefficient</i>	<i>Standard error</i>	<i>t</i>	<i>Degrees of freedom</i>	<i>p-value</i>
<i>Intercept, γ_{000}</i>	87.28	0.54	161.81	24	< .001
<i>Treatment, γ_{010}</i>	3.84	0.94	4.08	42	< .001
<i>Screenener, γ_{100}</i>	13.13	1.10	11.97	223	< .001
<i>Random effects</i>	<i>Standard deviation</i>	<i>Variance component</i>	<i>Degrees of freedom</i>	χ^2	<i>p-value</i>
<i>Student, e_{ijk}</i>	9.57	91.49			
<i>School, r_{0jk}</i>	0.17	0.03	24	26.03	.351
<i>Pair, u_{00k}</i>	1.65	2.73	24	45.43	.005

Note: A three-level hierarchical linear model and five multiply imputed datasets were used to estimate the statistics in this table. Student sample sizes: $n_{total} = 675$; $n_{intervention} = 414$, $n_{control} = 261$. School sample sizes: $n_{total} = 50$; $n_{intervention} = 25$, $n_{control} = 25$.

Source: Study data collected August 2008–May 2009.

Table M-3. Sensitivity analysis 2: without matched pairs

<i>Fixed effects model</i>	<i>Coefficient</i>	<i>Standard error</i>	<i>t</i>	<i>Degrees of freedom</i>	<i>p-value</i>
<i>Intercept, γ_{00}</i>	86.21	(0.45)	190.27	74	< .001
<i>Treatment, γ_{01}</i>	4.28	(0.92)	4.62	74	< .001
<i>Screenener, γ_{10}</i>	13.66	(0.89)	15.43	260	< .001
<i>Random effects</i>	<i>Standard deviation</i>	<i>Variance component</i>	<i>Degrees of freedom</i>	χ^2	<i>p-value</i>
<i>Student, e_{ij}</i>	(9.82)	96.40			
<i>School, r_{0j}</i>	(1.98)	3.94	74	117.01	.001

Note: A two-level hierarchical linear model and five multiply imputed datasets were used to estimate the statistics in this table. Student sample sizes: $n_{total} = 994$; $n_{intervention} = 615$, $n_{control} = 379$. School sample sizes: $n_{total} = 76$; $n_{intervention} = 38$, $n_{control} = 38$.

Source: Study data collected August 2008–May 2009.

Table M-4. Sensitivity analysis 3: without baseline covariate (screener)

<i>Fixed effects model</i>	<i>Coefficient</i>	<i>Standard error</i>	<i>t</i>	<i>Degrees of freedom</i>	<i>p-value</i>
<i>Intercept, γ_{000}</i>	86.3	(0.56)	155.10	37	< .001
<i>Treatment, γ_{010}</i>	4.16	(0.87)	4.78	70	< .001
<i>Random effects</i>	<i>Standard deviation</i>	<i>Variance component</i>	<i>Degrees of freedom</i>	χ^2	<i>p-value</i>
<i>Student, e_{ijk}</i>	(11.06)	122.42			
<i>School, r_{0jk}</i>	(0.22)	0.05	37	42.27	.253
<i>Pair, u_{00k}</i>	(2.19)	4.78	37	74.75	< .001

Note: A three-level hierarchical linear model and five multiply imputed datasets were used to estimate the statistics in this table. Student sample sizes: $n_{total} = 994$; $n_{intervention} = 615$, $n_{control} = 379$. School sample sizes: $n_{total} = 76$; $n_{intervention} = 38$, $n_{control} = 38$.

Source: Study data collected from August 2008–May 2009.

Table M-5. Sensitivity analysis 4: using cases with complete Test of Early Mathematics Ability–Third Edition (Ginsburg and Baroody 2003) scores only

<i>Fixed effects model</i>	<i>Coefficient</i>	<i>Standard error</i>	<i>t</i>	<i>Degrees of freedom</i>	<i>p-value</i>
<i>Intercept, γ_{000}</i>	86.53	(0.42)	207.40	37	< .001
<i>Treatment, γ_{010}</i>	3.82	(0.70)	5.45	74	< .001
<i>Screener, γ_{100}</i>	13.58	(0.87)	15.66	878	< .001
<i>Random effects</i>	<i>Standard deviation</i>	<i>Variance component</i>	<i>Degrees of freedom</i>	χ^2	<i>p-value</i>
<i>Student, e_{ijk}</i>	(9.64)	92.95			
<i>School, r_{0jk}</i>	(0.30)	0.09	37	35.49	> .500
<i>Pair, u_{00k}</i>	(1.38)	1.92	37	55.29	.027

Note: A three-level hierarchical linear model and five multiply imputed datasets were used to estimate the statistics in this table. Student sample sizes: $n_{total} = 881$; $n_{intervention} = 555$, $n_{control} = 326$. School sample sizes: $n_{total} = 76$; $n_{intervention} = 38$, $n_{control} = 38$.

Source: Study data collected August 2008–May 2009.

Table M-6. Sensitivity analysis 5: excluding students assigned to tutoring groups with students who were not part of the at-risk analytic sample

<i>Fixed effects model</i>	<i>Coefficient</i>	<i>Standard error</i>	<i>t</i>	<i>Degrees of freedom</i>	<i>p-value</i>
<i>Intercept, γ_{000}</i>	86.00	(0.50)	172.75	37	< .001
<i>Treatment, γ_{010}</i>	4.15	(0.85)	4.86	68	< .001
<i>Screener, γ_{100}</i>	13.76	(0.91)	15.18	264	< .001
<i>Random effects</i>	<i>Standard deviation</i>	<i>Variance component</i>	<i>Degrees of freedom</i>	χ^2	<i>p-value</i>
<i>Student, e_{ijk}</i>	(9.88)	97.61			
<i>School, σ_{jk}</i>	(1.02)	1.03	31	43.36	.070
<i>Pair, u_{00k}</i>	(1.65)	2.73	37	61.94	.006

Note: A three-level hierarchical linear model and five multiply imputed datasets were used to estimate the statistics in this table. Student sample sizes: $n_{total} = 970$; $n_{intervention} = 591$, $n_{control} = 379$. School sample sizes: $n_{total} = 70$; $n_{intervention} = 35$, $n_{control} = 35$.

Source: Study data collected August 2008–May 2009.

Table M-7. Sensitivity analysis 6: excluding school pairs with tutoring groups that included students who were not part of the at-risk analytic sample

<i>Fixed effects model</i>	<i>Coefficient</i>	<i>Standard error</i>	<i>t</i>	<i>Degrees of freedom</i>	<i>p-value</i>
<i>Intercept, γ_{000}</i>	86.25	(0.52)	167.18	28	< .001
<i>Treatment, γ_{010}</i>	4.08	(0.87)	4.683	56	< .001
<i>Screener, γ_{100}</i>	14.12	(0.96)	14.768	256	< .001
<i>Random effects</i>	<i>Standard deviation</i>	<i>Variance component</i>	<i>Degrees of freedom</i>	χ^2	<i>p-value</i>
<i>Student, e_{ijk}</i>	(9.891)	98.16			
<i>School, σ_{jk}</i>	(1.04)	1.08	28	40.46	.060
<i>Pair, u_{00k}</i>	(1.59)	2.53	28	48.79	.009

Note: A three-level hierarchical linear model and five multiply imputed datasets were used to estimate the statistics in this table. Student sample sizes: $n_{total} = 883$; $n_{intervention} = 579$, $n_{control} = 304$. School sample sizes: $n_{total} = 58$; $n_{intervention} = 29$, $n_{control} = 29$.

Source: Study data collected August 2008–May 2009.

Analyses reported in chapter 5

Table M-8. Exploratory 1: differential impact based on baseline mathematics proficiency

<i>Fixed effects model</i>	<i>Coefficient</i>	<i>Standard error</i>	<i>t</i>	<i>Degrees of freedom</i>	<i>p-value</i>
<i>Intercept, γ_{000}</i>	86.19	(0.49)	177.42	37	< .001
<i>Treatment, γ_{010}</i>	4.27	(0.82)	5.18	60	< .001
<i>Screener, γ_{100}</i>	13.48	(0.91)	14.81	74	< .001
<i>Interaction, γ_{000}</i>	1.06	(1.83)	0.58	74	0.564
<i>Random effects</i>	<i>Standard deviation</i>	<i>Variance component</i>	<i>Degrees of freedom</i>	χ^2	<i>p-value</i>
<i>Student, e_{ijk}</i>	(9.81)	96.21			
<i>School, r_{0jk}</i>	(0.91)	0.83	35	45.52	.110
<i>Interaction, r_{1jk}</i>	(0.84)	0.70	72	65.92	> .500
<i>Pair, u_{00k}</i>	(1.64)	2.69	37	61.83	.007

Note: A three-level hierarchical linear model and five multiply imputed datasets were used to estimate the statistics in this table. Student sample sizes: $n_{total} = 994$; $n_{intervention} = 615$, $n_{control} = 379$. School sample sizes: $n_{total} = 76$; $n_{intervention} = 38$, $n_{control} = 38$.

Source: Study data collected August 2008–May 2009.

Table M-9. Exploratory 2: effect on letter- and word-reading proficiency for students participating in *Number Rockets*, for cases with complete Woodcock Johnson–Third Edition Letter/Word (Woodcock, McGrew, and Mather 2001 subtest-reading scores only

<i>Fixed effects model</i>	<i>Coefficient</i>	<i>Standard error</i>	<i>t</i>	<i>Degrees of freedom</i>	<i>p-value</i>
<i>Intercept, γ_{000}</i>	97.94	(0.73)	133.33	37	< .001
<i>Treatment, γ_{010}</i>	0.12	(1.13)	0.11	74	.913
<i>Screener, γ_{100}</i>	11.99	(1.21)	9.87	283	< .001
<i>Random effects</i>	<i>Standard deviation</i>	<i>Variance component</i>	<i>Degrees of freedom</i>	χ^2	<i>p-value</i>
<i>Student, e_{ijk}</i>	(13.38)	179.12			
<i>School, r_{0jk}</i>	(1.85)	3.43	37	49.98	0.075
<i>Pair, u_{00k}</i>	(3.03)	9.15	37	76.86	< .001

Note: A three-level hierarchical linear model and five multiply imputed datasets were used to estimate the statistics in this table. Student sample sizes: $n_{total} = 892$; $n_{intervention} = 558$, $n_{control} = 334$. School sample sizes: $n_{total} = 76$; $n_{intervention} = 38$, $n_{control} = 38$.

Source: Study data collected August 2008–May 2009.

Table M-10. Exploratory 3: Relationship between implementation level and school-pair level impact of *Number Rockets*

<i>Fixed effects model</i>	<i>Coefficient</i>	<i>Standard error</i>	<i>t</i>	<i>Degrees of freedom</i>	<i>p-value</i>
<i>Intercept, γ_{000}</i>	86.18	(0.50)	171.776	36	< .001
<i>Treatment, γ_{010}</i>	4.25	(0.82)	5.155	60	< .001
<i>Screener, γ_{100}</i>	13.64	(0.90)	15.19	307	< .001
<i>Interaction, γ_{000}</i>	0.06	(0.10)	0.555	36	.582
<i>Random effects</i>	<i>Standard deviation</i>	<i>Variance component</i>	<i>Degrees of freedom</i>	χ^2	<i>p-value</i>
<i>Student, e_{ijk}</i>	(9.82)	96.45			
<i>School, r_{0jk}</i>	(0.63)	0.40	37	46.87	.128
<i>Pair, u_{00k}</i>	(1.72)	2.95	36	65.82	.002

Note: A three-level hierarchical linear model and five multiply imputed datasets were used to estimate the statistics in this table. Student sample sizes: $n_{total} = 994$; $n_{intervention} = 615$, $n_{control} = 379$. School sample sizes: $n_{total} = 76$; $n_{intervention} = 38$, $n_{control} = 38$.

Source: Study data collected August 2008–May 2009.

Table M-11. Exploratory 1 sensitivity analysis: effect of *Number Rockets* for lowest third of students at-risk for mathematics difficulties

<i>Fixed effects model</i>	<i>Coefficient</i>	<i>Standard error</i>	<i>t</i>	<i>Degrees of freedom</i>	<i>p-value</i>
<i>Intercept, γ_{000}</i>	80.07	(0.63)	127.45	37	< .001
<i>Treatment, γ_{010}</i>	3.87	(1.17)	3.30	66	.002
<i>Screener, γ_{100}</i>	17.27	(2.56)	6.73	328	< .001
<i>Random effects</i>	<i>Standard deviation</i>	<i>Variance component</i>	<i>Degrees of freedom</i>	χ^2	<i>p-value</i>
<i>Student, e_{ijk}</i>	(9.57)	91.50			
<i>School, r_{0jk}</i>	(0.19)	0.04	29	23.63	> .500
<i>Pair, u_{00k}</i>	(0.62)	0.38	37	38.71	.392

Note: A three-level hierarchical linear model and five multiply imputed datasets were used to estimate the statistics in this table. Student sample sizes: $n_{total} = 331$; $n_{intervention} = 217$, $n_{control} = 114$. School sample sizes: $n_{total} = 76$; $n_{intervention} = 38$, $n_{control} = 38$.

Source: Study data collected August 2008–May 2009.

Table M-12. Exploratory 1 sensitivity analysis: effect for middle third of students at-risk for mathematics difficulties

<i>Fixed effects model</i>	<i>Coefficient</i>	<i>Standard error</i>	<i>t</i>	<i>Degrees of freedom</i>	<i>p-value</i>
<i>Intercept, γ_{000}</i>	87.45	(0.76)	115.04	37	< .001
<i>Treatment, γ_{010}</i>	3.38	(1.30)	2.59	67	.012
<i>Screener, γ_{100}</i>	10.81	(4.85)	2.23	328	.027
<i>Random effects</i>	<i>Standard deviation</i>	<i>Variance component</i>	<i>Degrees of freedom</i>	χ^2	<i>p-value</i>
<i>Student, e_{ijk}</i>	(9.34)	87.32			
<i>School, r_{0jk}</i>	(1.04)	1.08	30	26.45	> .500
<i>Pair, u_{00k}</i>	(2.50)	6.24	37	63.02	.005

Note: A three-level hierarchical linear model and five multiply imputed datasets were used to estimate the statistics in this table. Student sample sizes: $n_{total} = 331$; $n_{intervention} = 193$, $n_{control} = 138$. School sample sizes: $n_{total} = 76$; $n_{intervention} = 38$, $n_{control} = 38$.

Source: Study data collected August 2008–May 2009.

Table M-13. Exploratory 1 sensitivity analysis: effect for highest third of students at-risk for mathematics difficulties

<i>Fixed effects model</i>	<i>Coefficient</i>	<i>Standard error</i>	<i>t</i>	<i>Degrees of freedom</i>	<i>p-value</i>
<i>Intercept, γ_{000}</i>	91.18	(0.62)	147.40	36	< .001
<i>Treatment, γ_{010}</i>	5.36	(1.33)	4.04	67	< .001
<i>Screener, γ_{100}</i>	13.17	(6.48)	2.03	329	0.043
<i>Random effects</i>	<i>Standard deviation</i>	<i>Variance component</i>	<i>Degrees of freedom</i>	χ^2	<i>p-value</i>
<i>Student, e_{ijk}</i>	(10.53)	110.93			
<i>School, r_{0jk}</i>	(0.21)	0.04	31	30.94	> .500
<i>Pair, u_{00k}</i>	(0.21)	0.04	36	33.02	> .500

Note: A three-level hierarchical linear model and five multiply imputed datasets were used to estimate the statistics in this table. Student sample sizes: $n_{total} = 332$; $n_{intervention} = 205$, $n_{control} = 127$. School sample sizes: $n_{total} = 76$; $n_{intervention} = 38$, $n_{control} = 38$.

Source: Study data collected August 2008–May 2009.

Table M-14. Exploratory 1 sensitivity analysis: using lowest third as the reference group

<i>Fixed effects model</i>	<i>Coefficient</i>	<i>Standard error</i>	<i>t</i>	<i>Degrees of freedom</i>	<i>p-value</i>
<i>Intercept, γ_{000}</i>	86.15	(0.50)	172.26	37	< .001
<i>Treatment, γ_{010}</i>	4.41	(0.81)	5.44	59	< .001
<i>Middle third, γ_{100}</i>	7.53	(0.88)	8.52	191	< .001
<i>Middle third*Treatment, γ_{110}</i>	-0.75	(1.66)	-0.45	988	.653
<i>Upper third, γ_{200}</i>	11.21	(0.86)	13.01	712	< .001
<i>Upper third*treatment, γ_{210}</i>	0.88	(1.73)	0.51	355	.611
<i>Random effects</i>	<i>Standard deviation</i>	<i>Variance component</i>	<i>Degrees of freedom</i>	χ^2	<i>p-value</i>
<i>Student, e_{ijk}</i>	(10.07)	101.33			
<i>School, r_{0jk}</i>	(0.43)	0.19	37	87.79	< .001
<i>Pair, u_{00k}</i>	(1.84)	3.40	37	69.73	.001

Note: A three-level hierarchical linear model and five multiply imputed datasets were used to estimate the statistics in this table. Student sample sizes: $n_{total} = 994$; $n_{intervention} = 615$, $n_{control} = 379$. School sample sizes: $n_{total} = 76$; $n_{intervention} = 38$, $n_{control} = 38$.

Source: Study data collected August 2008–May 2009.

Table M-15. Exploratory 1 sensitivity analysis: using highest third as the reference group

<i>Fixed effects model</i>	<i>Coefficient</i>	<i>Standard error</i>	<i>t</i>	<i>Degrees of freedom</i>	<i>p-value</i>
<i>Intercept, γ_{000}</i>	86.15	(0.50)	172.26	37	< .001
<i>Treatment, γ_{010}</i>	4.41	(0.81)	5.44	59	< .001
<i>Lower third, γ_{100}</i>	-11.21	(0.86)	-13.01	712	< .001
<i>Lower third*treatment, γ_{110}</i>	-0.88	(1.73)	-0.51	355	.611
<i>Middle third, γ_{200}</i>	-3.69	(0.82)	-4.49	988	< .001
<i>Middle third*treatment, γ_{210}</i>	-1.63	(1.81)	-0.90	91	.373
<i>Random effects</i>	<i>Standard deviation</i>	<i>Variance component</i>	<i>Degrees of freedom</i>	χ^2	<i>p-value</i>
<i>Student, e_{ijk}</i>	(10.07)	101.33			
<i>School, r_{0jk}</i>	(0.43)	0.19	37	87.79	< .001
<i>Pair, u_{00k}</i>	(1.84)	3.40	37	69.73	.001

Note: A three-level hierarchical linear model and five multiply imputed datasets were used to estimate the statistics in this table. Student sample sizes: $n_{total} = 994$; $n_{intervention} = 615$, $n_{control} = 379$. School sample sizes: $n_{total} = 76$; $n_{intervention} = 38$, $n_{control} = 38$.

Source: Study data collected August 2008–May 2009.

References

- Agodini, R., Harris, B., Atkins-Burnett, S., Heaviside, S., Novak, T., Murphy, R., et al. (2009). *Achievement effects of four early elementary school math curricula: Findings from first graders in 39 schools*. (NCEE 2009-4052). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved June 8, 2010, from <http://ies.ed.gov/ncee/pubs/20094052/pdf/20094052.pdf>
- Arnold, D. H., and Doctoroff, G. L. (2003). The early education of socioeconomically disadvantaged children. *Annual Review of Psychology*, 54, 517–545.
- Arkansas State Department of Education Special Education Unit. (2010). *Part B state performance plan 2005–2010*. Retrieved June 9, 2010, from http://arksped.k12.ar.us/documents/data_n_research/ar-spprev-2010b.pdf
- Baker, S., Gersten, R., and Lee, D. (2002). A synthesis of empirical research on teaching mathematics to low-achieving students. *The Elementary School Journal*, 103, 51–73.
- Baker, S., Gersten, G., Flojo, J., Katz, R., Chard, D. J., and Clarke, B. (2006). *Preventing math difficulties in young children: Focus on effective screening of early number sense delays*. (Pacific Institutes for Research Technical Report No. 0305). Eugene, OR: Pacific Institutes for Research.
- Barnes, A. C., and Harlacher, J. E. (2008). Clearing the confusion: Response-to-intervention as a set of principles. *Education and Treatment of Children*, 31(3), 417–431.
- Baroody, A. J., Li, X., and Lai, M. (2008). Toddlers' spontaneous attention to number. *Mathematical Thinking and Learning*, 10, 240–270.
- Berch, D. B. (2005). Making sense of number sense: Implications for children with mathematical disabilities. *Journal of Learning Disabilities*, 38(4), 333–339.
- Binet, A., and Simon, T. H. (1905/1916). *The development of intelligence in children*. (E.S. Kite, Trans.). Baltimore, MD: Williams and Wilkins Company.
- Bloom, H. S. (2005). *Learning more from social experiments*. New York: Sage.
- Bloom, H. S., Richburg-Hayes, L. and Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, 29(1), 30–59.
- Booth, J. L., and Siegler, R. S. (2008). Numerical representations influence arithmetic learning. *Child Development*, 79(4), 1016–1031.
- Bryant, D. P., Bryant, B. R., Gersten, R., Scammacca, N., and Chavez, 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 (1), 20–32.
- Burghardt, J., Deke, J., Kisker, E., Puma, M., and Schochet, P. (2009). *Regional educational laboratory rigorous applied research studies: Frequently asked analysis questions*. Washington, DC: U.S. Department of Education, Institute for Education Sciences.
- Butler, F. M., Miller, S. P., Crehan, K., Babbitt, B., and Pierce, T. (2003). Fraction instruction for students with mathematics disabilities: Comparing two teaching sequences. *Learning Disabilities Research and Practice*, 18, 99–111.
- Case, R., Okamoto, Y., Griffin, S., McKeough, A., Bleiker, C., Henderson, B., et al. (1996). The role of central conceptual structures in the development of children's thought. *Monographs of the Society for Research in Child Development*, 61, 1–2.
- Clarke, B., Baker, S., Chard, D., and Otterstedt, J. (2006). *Developing and validating measures of number sense to identify students at-risk for mathematics disabilities*. (Pacific Institutes for Research Technical Report No. 0307). Eugene, OR: Pacific Institutes for Research.
- Clarke, B., Gersten, R., and Newman-Gonchar, R. (2010). RTI in mathematics: Beginnings of a knowledge base. In S. Vaughn and T. A. Glover (Eds.), *The promise of response to intervention: Evaluating the current science and practice* (pp. 187–203). New York: Guilford Press.
- Clements, D. H., and McMillen, S. (1996). Rethinking “concrete” manipulatives. *Teaching Children Mathematics*, 2(5), 270–279.
- Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., et al. (2010). Selecting at-risk first grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology*, 102(2), 327–340.
- Crehan, K. D. (2005). Review of the Test of Early Mathematics Ability—Third Edition. In Spies R. A. and Plake, B. S. (Eds.), *The sixteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements. Retrieved March 8, 2011, from the Mental Measurements Yearbook with Tests in Print database.
- Duncan, G., Dowsett, C., Claessens, A., Magnuson, K., Huston, A., Klebanov, et al. (2007). School readiness and later achievement. *Developmental Psychology*, 43(6), 1428–1446.
- Fuchs, L., Compton, D., Fuchs, D., Paulsen, K., Bryant, J., and Hamlett, C. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology*, 97, 493–513.

- Fuchs, L. S., Fuchs, D., Bryant, J. D., Hamlett, C. L., and Seethaler, P. M. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Exceptional Children*, 73(3), 311–330.
- Fuchs, L., Fuchs, D., Craddock, C., Hollenbeck, K., Hamlett, C., and Schatschneider, C. (2008). Effects of small-group tutoring with and without validated classroom instruction on at-risk students' math problem solving: Are two tiers of prevention better than one? *Journal of Educational Psychology*, 100(3), 491–509.
- Fuchs, L., Hamlett, C., and Fuchs, D. (1990). *Curriculum-based math computation and concepts/applications*. Nashville, TN: Vanderbilt University.
- Fuchs, L. S., Hamlett, C. L., and Powell, S. R. (2003). Fact fluency assessment. (Available from L. S. Fuchs, 328 Peabody, Vanderbilt University, Nashville, TN 37203).
- Geary, D. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin*, 114, 345–362.
- Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J., et al. (2009). *Assisting students struggling with mathematics: Response to Intervention (RtI) for elementary and middle schools* (NCEE 2009-4060). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved June 8, 2010, from <http://ies.ed.gov/ncee/wwc/publications/practiceguides/>
- Gersten, R., Compton, D., Connor, C. M., Dimino, J., Santoro, L., Linan-Thompson, S., et al. (2008). *Assisting students struggling with reading: Response to Intervention and multi-tier intervention for reading in the primary grades. A practice guide*. (NCEE 2009-4045). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved June 16, 2010, from <http://ies.ed.gov/ncee/wwc/publications/practiceguides/>
- Gersten, R., Jordan, N. C., and Flojo, J. R. (2005). Early identification and interventions for students with math difficulties. *Journal of Learning Disabilities*, 38(4), 293–304.
- Ginsburg, H., and Baroody, A. (2003). *Test of Early Mathematics Ability—Third Edition*. Austin, TX: Pro-Ed.
- Glover, T., and Diperna, J. (2007). Service delivery for response to intervention: Core components and directions for future research. *School Psychology Review*, 36(4), 526–540.
- Good, R. H., and Kaminski, R. A. (Eds.). (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

- Graham, T. A., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 8, 206-213.
- Griffin, E. A. (1997, April). *The role of children's social skills in achievement at kindergarten entry and beyond*. Poster session presented at the Biennial Meeting of the Society for Research in Child Development, Washington, DC.
- Hanich, L. B., Jordan, N. C., Kaplan, D., and Dick, J. (2001). Performance across different areas of mathematical cognition in children with learning difficulties. *Journal of Educational Psychology*, 93(3), 615–629.
- Harcourt Assessment, Inc. (2004). *Stanford Achievement Test Series, Tenth Edition*. San Antonio, TX: Pearson Education Measurement, Inc.
- Harcourt Brace Educational Measurement. (1996). *Stanford Early School Achievement Test-Fourth Edition [SESAT-2]*. Orlando, FL: Harcourt Brace Educational Measurement.
- Henry, V., and Brown, R. (2008). First-grade basic facts. *Journal for Research in Mathematics Education*, 39, 153–183.
- Hill, C. J., Bloom, H. S., Black, A. R., and Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, 2(3), 172–177.
- Hojnoski, R. L., Silberglitt, B., and Floyd, R. G. (2009). Sensitivity to growth over time of the preschool numeracy indicators with a sample of preschoolers in Head Start. *School Psychology Review*, 38(3), 402–418.
- Hruz, T. (July, 2002). *Wisconsin Policy Research Institute Report: The growth of special education in Wisconsin*. Retrieved October 12, 2010, from <http://www.wpri.org/Reports/Volume15/Vol15no5/Vol15no5summary.pdf>
- Individuals with Disabilities Education Improvement Act, 20 U.S.C. § 1400 (2004).
- Institute of Education Sciences. (2008). *WWC Procedures and Standards Handbook: Appendix B - Effect Size Computations (Version 2.0 – December 2008)*. Retrieved October 21, 2010, from <http://ies.ed.gov/ncee/wwc/help/idocviewer/Doc.aspx?docId=19&tocId=8>
- Institute of Education Sciences. (2011). *WWC Procedures and Standards Handbook*. Retrieved January 26, 2011, from <http://ies.ed.gov/ncee/wwc/references/idocviewer/doc.aspx?docId=19&tocId=7>
- Jordan, N., and Hanich, L. (2000). Mathematical thinking in second grade children with different forms of LD. *Journal of Learning Disabilities*, 23(6), 567–578.

- Jordan, N. C., Hanich, L. B., and Kaplan, D. (2003). A longitudinal study of mathematical competencies in children with mathematics difficulties with and without co-morbid reading difficulties. *Child Development*, 74, 834–850.
- Jordan, N. C., Kaplan, D., and Hanich, L.B. (2002). Achievement growth in children with learning difficulties in mathematics: Findings of a two-year longitudinal study. *Journal of Educational Psychology*, 94, 586–597.
- Jordan, N. C., Kaplan, D., Oláh, L. N., and Locuniak, M. N. (2006). Number sense growth in kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child Development*, 77(1), 153–175.
- Jordan, N. C., Kaplan, D., Locuniak, M. N., and Ramineni, C. (2007). Predicting first-grade math achievement from developmental number sense trajectories. *Learning Disabilities Research and Practice*, 22(1), 36–46.
- Kalchman, M., Moss, J., and Case, R. (2001). Psychological models for the development of mathematical understanding: Rational numbers and functions. In S. Carver and D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 1–38), Mahwah, NJ: Erlbaum.
- Louisiana Department of Education. (2009). *RTI Policy approved by BESE, June 2009*. Retrieved June 8, 2010, from <http://www.louisianaschools.net/lde/uploads/15858.pdf>
- Mazzocco, M. M., and Myers, G. F. (2003). Complexities in identifying and defining mathematics learning disability in the primary school-age years. *Annals of Dyslexia*, 53, 218–253.
- Methe, S. A., Hintze, J. M., and Floyd, R. G. (2008). Validation and decision accuracy of early numeracy skill indicators. *School Psychology Review*, 37(3), 359–373.
- Monsaas, J. A. (2005). Review of the Test of Early Mathematics Ability—Third Edition. In R. A. Spies and B.S. Plake. (Eds.), *The sixteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements. Retrieved March 8, 2011, from the Mental Measurements Yearbook with Tests in Print database.
- Morgan, P., Farkas, G., and Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities*, 42(4), 306–321.
- National Center for Education Statistics. (n.d.). *Common Core of Data* (2008/09). Retrieved September 26, 2010, from <http://nces.ed.gov/ccd/>
- National Center for Education Statistics. (2004). *2003–2004 School Staffing Survey*. Retrieved July 16, 2010, from http://nces.ed.gov/surveys/sass/tables_list.asp

- National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence*. Reston, VA: National Council of Teachers of Mathematics.
- National Council of Teachers of Mathematics. (2009). *2009 NCTM legislative platform*. Retrieved March 5, 2010, from http://www.nctm.org/uploadedFiles/Research_Issues_and_News-Section_Navigation/Legislation/2009_Leg_Platform.pdf
- National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Retrieved August 14, 2009, from <http://www.ed.gov/about/bdscomm/list/mathpanel/report/final-report.pdf>
- National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Mathematics Learning Study Committee, J. Kilpatrick, J. Swafford, and B. Findell (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- National Research Council. (2009). *Mathematics learning in early childhood: Paths toward excellence and equity*. Committee on Early Childhood Mathematics, Christopher T. Cross, Taniesha A. Woods, and Heidi Schweingruber (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- New Mexico Public Education Department. (2009). *Understanding and implementing the response to intervention (RtI) framework in New Mexico: A quick guide*. Retrieved June 8, 2010, from <http://www.ped.state.nm.us/RtI/dl09/Understanding%20Response%20to%20Inter.pdf>
- Newman-Gonchar, R., Clarke, B., and Gersten, R. (2009). *A summary of nine key studies: Multitier intervention and response to interventions for struggling students in mathematics*. Portsmouth, NH: RMC Research Corporation, Center on Instruction.
- O'Conner, R. E., Harty, K. R., and Fulmer, D. (2005). Tiers of intervention in kindergarten through third grade. *Journal of Learning Disabilities*, 38(6), 532–538.
- Oklahoma State Department of Education. (2007). *Policies and procedures for special education in Oklahoma*. Retrieved June 9, 2010, from http://sde.state.ok.us/Curriculum/SpecEd/pdf/Compliance/Policies_Procedures.pdf
- Paulsen, K., and Fuchs, L. (2005). *First-grade small group tutoring to prevent math difficulty: Volume I—Scripts*. Unpublished document.
- Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., and Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal*, 45, 365–397.
- Raudenbush, S.W., Bryk, A.S., & Congdon, R. (2006). HLM 6.02 for Windows [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

- Raudenbush, S. W., Martinez, A., and Spybrook, J. (2005). *Strategies for improving precision in group-randomized experiments*. New York: William T. Grant Foundation.
- Reid, D. K., Hresko, W. P., and Hammill, D. D. (1989). *Test of Early Reading—Second Edition*. Austin, TX: Pro-Ed.
- Resnick, L. B. (1982). Syntax and semantics in learning to subtract. In T. P. Carpenter, J. M. Moser, and T. A. Romberg (Eds.), *Addition and subtraction: A cognitive perspective* (pp. 136-155). Hillsdale, NJ: LEA.
- Roza, M., Guin, K., and Davis, T. (2008). *What is the sum of the parts? How federal, state, and district funding streams confound efforts to address different student types*. Seattle, WA: Center on Reinventing Public Education, University of Washington.
- Rubin, D. (1987). *Multiple imputation for non-response in surveys*. New York: Wiley.
- Sattler, J. M., and Hoge, R. D. (2006). *Assessment of children: Behavioral, social, and clinical foundations* (5th ed.). La Mesa, CA: Jerome M. Sattler, Publisher, Inc.
- Schochet, P. (2005). *Statistical power analysis for random assignment evaluations of education programs*. Retrieved June 10, 2010, from <http://www.mathematica-mpr.com/publications/>
- Speece, D. L., Ritchey, K. D., Cooper, D. H., Roth, F. P., and Schatschneider, C. (2004). Growth in early reading skills from kindergarten to third grade. *Contemporary Educational Psychology*, 29(3), 312–332
- Texas Education Agency. (2008). *2008–2009 response to intervention guidance*. Retrieved June 8, 2010, from <http://ritter.tea.state.tx.us/curriculum/RtI/RtIGuidanceDocument.pdf>
- Tilly, D. (2003, December). *Heartland Area Education Agency's evolution from four to three tiers: Our journey—Our results*. Paper presented at the National Research Center on Learning Disabilities Responsiveness-to-Intervention Symposium, Kansas City, MO.
- Tomlinson, C. A. (1999). *The differentiated classroom: Responding to the needs of all learners*. Alexandria, VA: Association for Supervision and Curriculum Development.
- U.S. Department of Education. (n.d.). *NAEP data explorer*. Retrieved August 17, 2010, from <http://nces.ed.gov/nationsreportcard/naepdata/>
- Vaughn, S., and Fuchs, L. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning Disabilities Research and Practice*, 18, 137–146.

- Vaughn, S., Linan-Thompson, S., and Hickman, P. (2003). Means of identifying students with reading/learning disability. *Council for Exceptional Children*, 69(4), 391–409.
- Vaughn, S., Moody, S. W., and Schumm, J. S. (1998). Broken promises: Reading instruction in the resource room. *Exceptional Children*, 64(2), 211–225.
- Wechsler, D. (2003). *Wechsler Intelligence Scale for Children*®—Forth Edition. San Antonio, TX: The Psychological Corporation.
- Woodcock, R. W., and Johnson, M. B. (1990). *WJ-R tests of achievement: Examiner's manual*. Allen, TX: DLM Teaching Resources.
- Woodcock, R., McGrew, K., and Mather, N. (2001). *Woodcock-Johnson III*. Itasca, IL: Riverside Publishing.
- Xu, Z., Hannaway, J., and D'Souza, S. (2009). *Student transience in North Carolina: The effects of school mobility on student outcomes using longitudinal data*. (CALDER Working Paper No. 22). Washington, DC: The Urban Institute.

