EVALUATING MATH RECOVERY: THE IMPACT OF IMPLEMENTATION FIDELITY ON STUDENT OUTCOMES

By

Charles Munter

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Teaching and Learning

August, 2010

Nashville, Tennessee

Approved:

Paul Cobb

Thomas Smith

David Cordray

Rich Lehrer

ACKNOWLEDGEMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305B070554 and R305B040110 to Vanderbilt University. The opinions expressed are those of the author and do not necessarily represent the views of the U.S. Department of Education.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

One of the purposes of education research—and one that has been increasingly stressed in recent years with the enactment of the Education Science Reform Act of 2002 and the establishment of the Institute of Education Sciences (IES)—is to develop and rigorously evaluate programs to assess whether they are effective in supporting students' learning and achievement. This research agenda includes an emphasis on measuring implementation fidelity and linking those measures to program impacts (U.S. Department of Education, 2006). Claims of treatment (in)effectiveness may be unwarranted and invalid unless the degree to which programs are implemented as intended by their developers is defined and assessed. Indeed, direct assessments of implementation fidelity are necessary for making causal claims about the relationship between the components of an intervention and the outcomes of an evaluation (Bickman, 1987; Lipsey, 1993). Ideally, investigators go a step further and incorporate an index of the extent to which the intervention was faithfully implemented in their analyses of intervention outcomes (Nelson, Cordray, Hulleman, Darrow, & Sommer, 2010; O'Donnell, 2008).

But to date, analyses linking FOI indices to evaluation outcomes are not common in the literature. Despite the repeated arguments for the importance of assessing implementation fidelity, recent reviews of evaluations of school-based interventions have revealed that researchers typically do not assess fidelity of implementation of the programs they evaluate, let alone link fidelity indices to dependent variables (Darrow,

2009; Dusenbury, Brannigan, Falco, & Hansen, 2003; O'Donnell, 2008). This is especially true with respect to evaluations of unscripted interventions, where measuring fidelity first requires the identification and operationalization of complex, subtle facets of the intervention (Cordray & Pion, 2006).

In this dissertation, I report an analysis of the relationship between student outcomes and fidelity of implementation (FOI) of Math Recovery, an unscripted, pullout, tutoring program intended to increase the mathematics achievement of low-performing first graders. The work I describe was conducted as part of a larger evaluation study of MR. Two research questions guided the conduct and analysis of the larger evaluation study: 1) Does participation in MR raise the mathematics achievement of low performing first-grade students? 2) If so, do participating students maintain the gains made in first grade through the end of second grade? The work documented in this report was driven by a third question: 3) What is the relation between FOI of MR and student outcomes at the end of the school year in which students received tutoring? The relevance of the fidelity study stems from three broad goals of evaluation research: 1) assessing a program's potential for successful scale-up, 2) theory-testing, and 3) program improvement—none of which has previously been directly addressed with respect to the Math Recovery program.

Regarding the first goal, O'Donnell (2008) argued that assessments of FOI have implications for scale-up, which she defined as "the deliberate expansion of an externally developed program that has been previously proven efficacious in one or a small number of school settings to many settings" (p. 42). Figure 1 depicts the scaling-up decision-making process she described, which is based on two measures of an initial effectiveness

study: outcomes and fidelity. O'Donnell argued that if outcomes (e.g., student achievement) and fidelity are both high—indicating that the program can be implemented with fidelity and that it produced positive effects—the program can and should be adopted at a wide-scale. If, during the initial effectiveness trial, a program is not implemented with high fidelity, the results cannot be attributed to the prescribed program model, and therefore the program should not (yet) be taken to scale. As will be described below, the larger evaluation study in which the fidelity study was conducted found that the Math Recovery intervention failed to produce positive, lasting effects on student achievement. In the analyses reported in this dissertation, I investigate the relationship between fidelity of implementation and student outcomes to further investigate the reasons for this finding, which leads to a second goal of evaluation research.

Positive outcomes (high achievement)	Do not scale up	Scale up
Negative outcomes (low achievement)	Do not scale up	Do not scale up
	Low fidelity	High fidelity

Figure 1. Scale-up and fidelity during an effectiveness study (O'Donnell, 2008, p. 43)

Regarding the second area of relevance, Bickman (1987) defined program theory as "the construction of a plausible and sensible model of how a program is supposed to work" (p. 5), and argued that "[t]he nature of the generalizability process requires not only that the nature of the program be explicated but also the nature of the theory underlying the program be explicated" (p. 9). Bickman suggested that designing

evaluations that are guided by clear program theories is useful in a number of ways, including (a) identifying the problem that an intervention is intended to address; (b) developing or choosing appropriate measures; (c) drawing attention to intervening variables that specify causal linkages between various processes or subgoals and the overall goals of a program; and (d) more accurately discriminating between theory failure, program failure, and implementation failure. Well-conducted assessments of FOI inherently attend to program theory in the ways described by Bickman, thereby providing opportunities to test program developers' theory of how a program is expected to work by examining the impact of key components of a program on study outcomes (e.g., student achievement) and more thoroughly elaborating causal relationships.

Finally, to the extent that a fidelity study uncovers either program components that do not contribute (whether directly or indirectly) to positive outcomes or *non*-program components that *do* contribute to positive outcomes, the results can inform ongoing program refinement (Dane & Schneider, 1998).

The Math Recovery (MR) program is a useful case in which to address all three of these goals: determining scale-up potential, theory testing, and informing program improvement. First, the larger study in which I conducted the fidelity study was a relatively small-scale evaluation (i.e., 20 schools) of MR's effectiveness. As stated above, the evaluation found that the intervention did not produce lasting, positive effects on student achievement. These results fall into the bottom row of O'Donnell's (2008) 2x2 scale-up matrix (Figure 1), indicating that MR should therefore not be scaled up. The fidelity study I have conducted helps identify the correct column in the scale-up matrix, providing evidence as to why the program should not be scaled up (i.e., because it is not

effective, or because it was not successfully implemented).

Second, because it originated from research that investigated the process of young children's arithmetical learning (Steffe, von Glasersfeld, Richards, & Cobb, 1983; Steffe, Cobb, & von Glasersfeld, 1988), MR's underlying theory has a fairly well articulated history. Additionally, the program is guided by a set of explicit core principles so that, even if their operational definitions are largely implicit, a set of core program components for achieving its intended effects could be identified and their relationship to study outcomes tested.

Third, producing feedback for informing future program refinement is a reasonable goal with respect to MR because, although the program has served more than 3000 students in at least 19 states in the U.S. since 1999, it is a relatively new program. Developers (as opposed to representatives from publishing companies, for example) are still very much involved in program promotion and implementation. Therefore, the evaluation team has considerable access to MR developers at a time in the program's history at which it has not been implemented at a large scale, or institutionalized to an irreversible degree.

Summary of the Study

I conducted the fidelity study within a two-year randomized controlled trial of Math Recovery. The program consists of three primary components: 1) tutor training, 2) student identification and assessment and 3) one-to-one tutoring. It is the second and third of these to which the fidelity assessment pertained primarily, because it is in these

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

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

The larger evaluation study was carried out in 20 elementary schools (five urban, ten suburban and five rural), representing five districts in two states. Eighteen teachers were recruited to receive training and participate as MR tutors from the participating districts, with two of the tutors each serving two schools. In each year (2007-08 and 2008-09 academic years), 17 to 36 students deemed eligible (based on an initial MR screening) from each of 20 schools were randomly assigned to one of three tutoring cohorts or to the "wait list" for MR. The cohorts, consisting of three students each, were

staggered across different start dates (i.e., Cohort A—September, B—December, C—March). In both years, students on the randomly ordered waiting list were selected to join an MR tutoring cohort if an assigned participant left the school or was deemed "ineligible" due to a special education placement. The number of study participants totaled 517 in Year 1 and 510 in Year 2, of whom 172 received tutoring in Year 1 and 171 received tutoring in Year 2.

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

The fidelity assessment was guided by what we, in collaboration with program developers, determined were the core implementation components. Using the criteria for FOI identified by Dane and Schneider (1998), these core components comprised expectations pertaining to: (a) exposure and duration (i.e., number and length of tutoring lessons, time spent on strategy-based activities); (b) adherence (i.e., administering the MR initial assessment correctly, employing MR's "Learning Framework" to accurately diagnose students' thinking, assign profiles, and choose activities that align with the student's profile); and (c) quality of delivery. The last of these included (i) necessary and unique aspects of Math Recovery tutoring as compared to typical tutoring: the tutor's

ongoing assessment of the child's thinking and strategies (both reflective assessment between tutoring sessions and in-the-moment assessment), and the tutor's efforts to provide instruction within the child's zone of proximal development; (ii) necessary but not unique aspects: the nature of tutors' instruction (e.g., providing sufficient wait time after posing tasks); and (iii) prohibited behaviors (e.g., eliciting behaviors from the student or directly demonstrating methods for solving a problem). Additionally, our coding instruments included (iv) aspects of mathematics instruction identified in recent research on mathematics teaching as high quality forms of practice but that are (at least implicitly) prohibited by the MR tutoring model—instances of 'positive infidelity' (Cordray & Hulleman, 2009).

For the fidelity assessment, I randomly selected one student from each of the 6 cohorts across the two years of the study from each of the 18 tutors, for a total of 107 students (one tutor had only five cohorts due to a maternity leave). In line with the MR program model, nearly all tutoring sessions were video-recorded to aid tutor reflection and planning. These video recordings comprised the bulk of the data set I used to conduct the fidelity study. The initial assessment conducted at the beginning of every MR tutoring cycle and randomly selected 12 tutoring sessions were coded for each of the 107 students.

Five people, with experience in either elementary mathematics teaching, video coding, or both, were hired to code the fidelity data. They were first trained in MR tutoring (by the same individuals who trained the study tutors) and then in using the fidelity coding instruments. These training sessions were followed by four weeks of completely independent coding for which percent agreement was determined until an adequate level of agreement was reached consistently.

Throughout the final phase, another member of the research team and I met weekly with the coders to further refine, define and operationalize the aspects of MR that they were attempting to code. However, after four weeks of refinement work, agreement percentages plateaued at an inadequate level, largely due to differences in how coders 'chunked' the lessons they were coding (e.g., was it one big task, or two small tasks?) Therefore, we identified a central aspect of the MR Instructional Framework about which coders' structural decisions were consistently in agreement and for which all codes would remain relevant: coders coded only those parts of selected lessons in which tutors engaged students in strategy-based activities. After limiting our within-lesson fidelity assessment to such excerpts, coders were able to achieve an observer-agreement percentage of at least 80% on all classes of codes combined during the final training/refinement phase. Overall, the reliability of the coding instruments was high. For a majority of codes, disaggregated agreement rates were at least 0.70. For a few codes, however, agreement rates were not sufficiently high to include the data in my analyses.

I performed two tests to assess the validity of the instruments. First, video recordings of expert MR tutors—those used for MR training—were coded using the fidelity coding scheme described above. The scores of the MR expert tutors were generally high, suggesting that our instruments helped identify high-quality MR tutoring. Second, a subset of the fidelity video data, representing the full range of levels of fidelity to MR's expectations for tutoring in the study population, were submitted to 12 MR experts, all of whom had at least five years of experience with MR. These individuals remained blind to our coding instruments and other experts' ratings as they ranked and categorized (excellent, good, fair, or poor) video recordings of 1) the MR initial

assessment, 2) full lessons, and 3) lesson excerpts pertaining to only strategy-based activities.

The results of this second test of validity suggested that agreement among MR experts is not strong. For example, for pairs of MR experts, the mean Spearman rank correlation full lessons was only 0.24. Still, the rank correlation between experts' average rankings of full lessons (which I determined by calculating an average ranking for each video recording) and the rankings determined by fidelity scores was 0.43, suggesting a modest level of agreement between the results of our instruments and experts' assessments for full lessons.

My analyses were driven by 7 research questions: 1) Did MR tutors implement the program with fidelity? 2) To what extent is greater fidelity of implementation of MR's initial assessment associated with correct assignment of students' Learning Framework profiles at the outset of tutoring? 3) To what extent is greater fidelity of implementation of MR's structural aspects (i.e., aspects related to exposure, duration and adherence) associated with greater student outcomes? 4) To what extent is greater fidelity of implementation of MR's process aspects (i.e., aspects related to quality of delivery and participant responsiveness) associated with greater student outcomes? 5) To what extent is higher frequency of non-MR aspects of tutoring (i.e., positive infidelity) associated with greater student outcomes? 6) To what extent are all aspects of fidelity of implementation combined associated with greater student outcomes? 7) To what extent does student responsiveness, as measured by gains on the MR initial assessment, mediate the effect of tutoring on external mathematics assessments?

To assess the extent to which the MR intervention was implemented with fidelity (i.e., met model expectations), I calculated descriptive statistics and, where possible, compared them to levels of fidelity observed in the expert training video recordings. The results suggested that, overall, fidelity of implementation was inconsistent. On average, tutors met expectations with respect to the administration of the initial assessment; video-recording all sessions with students; length of tutoring sessions; posing tasks to students that are neither too easy nor too difficult; adjusting the difficulty of tasks that, when initially posed, were too difficult; and the "nature of instruction" (rates of providing sufficient wait time and refraining from eliciting student behaviors and demonstrating methods).

However, a number of aspects of the MR intervention were implemented with questionable levels of fidelity, including: initial diagnoses of students' profiles; total number of lessons provided to students; average amount of time per lesson spent on strategy-based activities; and tutors' uses of particular moves endorsed (e.g., asking students to check their solutions, or soliciting students' strategies) or prohibited (e.g., eliciting student behaviors) by the MR tutoring model.

To answer my other research questions, including testing the validity of components of the MR program theory and the relationship of FOI to student outcomes, I employed two-level hierarchical linear models (to account for the clustering of students within tutors) and variables that I constructed using only those fidelity indicators that were coded with adequate reliability. First, I assessed the extent to which tutors' adherence to the MR initial assessment protocol was associated with rates of accuracy in

assigning student profiles at the outset of tutoring. This analysis identified no significant relationships.

Then, I examined the relationship between study outcomes at the end of the year in which students were tutored and aspects of MR pertaining to exposure/duration and quality of delivery (Dane & Schneider, 1998). These analyses identified significant relationships between outcomes and a number of fidelity variables. For example, the proportion of time spent on strategy-based activities significantly predicted scores on the MR initial assessment and both the Math Fluency and Math Reasoning subtests of the Woodcock-Johnson. Specifically, the main effect sizes for the WJIII outcomes equate to a mere 2-4 additional minutes of strategy-based activities per lesson. The ratio of tutors' rates of eliciting student behaviors to rates of soliciting students' strategies were predictive of the same three outcomes. For example, on the Math Reasoning subtest, the effect size found in the main analysis equates to a change from eliciting behaviors 2.8 times as often soliciting strategies, to soliciting student strategies equally as often as eliciting behaviors (ratio = 1), or roughly one standard deviation shift.

I also examined the relationship between instances of 'positive infidelity' and outcomes. For Math Reasoning, the impact of the use of such practices was large. The results suggest that the main effect size for that outcome is equivalent to just a four percent increase in positive infidelity moves.

Finally, I performed a mediation analysis to test the soundness of a particular component of the MR program theory: that the effect of MR tutoring on students' mathematics achievement is mediated by an increase in the sophistication of students' arithmetical strategies. For this analysis I used the MR initial assessment as a measure of

increases in such strategies—or 'participant responsiveness,' in terms of Dane and Schneider's (1998) criteria for FOI. The results of my analysis suggest that this component of the MR model is valid. MR tutoring had a significant impact on gains in MR initial assessment scores. And, when I regressed more distal outcomes on both MR initial assessment gains and a dummy variable indicating participation in tutoring, the effect of treatment found in the main analysis was reduced to a non-significant level. This suggests that increases in strategy sophistication (i.e., gains in the MR initial assessment) indeed mediate the impact of tutoring on mathematics achievement.

The results of my fidelity assessment have a number of implications—both for interpreting the results of the main effects analysis in terms of the fidelity findings, and for informing potential improvement of the MR intervention. My analyses revealed that the intervention was implemented with inconsistent levels of fidelity. This suggests that the impact of the intervention on student outcomes could potentially have been significantly larger had levels of FOI been higher. But it also suggests that the feasibility of implementing the MR intervention at a wide scale in natural settings is likely lower for some aspects of the MR program than for others, and therefore the means by which MR supports implementation should be examined. For example, a key aspect of the tutor's role is to use MR's Learning and Instructional Frameworks to diagnose students' thinking and to identify appropriate types of tasks to pose during tutoring sessions. My analysis of tutors' accuracy in assigning profiles at the outset of tutoring—when they have just completed a protocol designed specifically to support tutors in making such judgments—found that tutors assigned accurate profiles only two-thirds of the time. This suggests that MR's current forms of support for tutor learning are likely insufficient, and that tutors

could potentially benefit from additional, follow-up training in using the MR Frameworks.

In this dissertation, I frame the assessment of the fidelity of Math Recovery implementation as a case for examining the feasibility of conducting fidelity studies of unscripted interventions in general—an endeavor that has, heretofore, not been documented in the literature. Math Recovery is a clear example of an unscripted intervention, in that tutors' decisions are guided by their ongoing assessments of students' thinking, rather than prescribed steps to be taken to enact the program. Tutors are expected to support children in constructing increasingly sophisticated arithmetical strategies by continually adjusting instruction to their current understandings of number. As a consequence, assessing fidelity of this program is not as simple as monitoring adherence to a script, but requires determining the extent to which tutors continually adjust instruction to a child's current level of mathematical reasoning.

Additionally, the case of MR fidelity is useful from a practical, analytic standpoint, in that tutors enacted the program with varying degrees of fidelity to the MR model. As Nelson et al. (2010) pointed out, an intervention implemented with consistently high fidelity will have insufficient variation in fidelity indices for examining relationships between FOI and outcomes. But this was not the case in the evaluation of MR. Likely due to the inherent complexity of MR tutoring, the fidelity assessment resulted in a distribution of indices of fidelity across tutors—variation that could be leveraged in examining which components of the MR model matter most and why.

The structure of the dissertation is as follows: In chapter 2, I provide a conceptualization of fidelity of program implementation, including a description of the

emerging consensus on the important aspects of fidelity to be assessed and previous ways of operationalizing those elements. Chapter 3 begins with a description of Math Recovery and the program evaluation in which the fidelity study was being conducted. Then, I detail the process of developing and testing instruments for assessing implementation fidelity of Math Recovery. In doing so, I provide a concrete account of how recent conceptions and standards of fidelity assessment (as described in chapter 2) were applied to the FOI assessment of MR, including (a) identifying the intervention's program theory and core components; (b) creating operational definitions of the intervention's core components; (c) developing coding instruments; (d) hiring and training coders; (e) instituting a sampling frame sufficient for generalizing fidelity findings to the study population; and (f) determining the reliability and validity of the instruments. Chapter 3 concludes with a description of my methods for incorporating the results of the fidelity assessment into the evaluation analyses. I begin with a description of the main effects analysis, followed by descriptions of the variables I used in the fidelity analysis, my research questions, and the models I employed to answer those questions. In chapter 4, I report the findings of my analyses. I conclude in chapter 5 with a discussion of the results, highlighting key aspects of the work that were particularly helpful in assessing FOI of MR in order to illustrate how these steps might be accomplished in fidelity studies of other unscripted interventions.

CHAPTER 2

CONCEPTUALIZING FIDELITY OF IMPLEMENTATION

The synthesis in this chapter is intended to provide a description of both the emerging consensus with respect to the conceptualization and role of fidelity of implementation, and the framework from which I approached the work that I report in this dissertation. A number of reviews have recently been conducted on fidelity of implementation studies within evaluations in the fields of education, mental health, social services, prevention research and related fields (Dane & Schneider, 1998; Darrow, 2009; Dusenbury et al., 2003; Mowbray, Holter, Teague, & Bybee, 2003; O'Donnell, 2008). Table 1 provides a summary of these five reviews. In all cases, the authors argued for the inclusion of implementation fidelity studies as a standard component of evaluation research. In doing so, they attempted to push their respective fields toward both a shared conceptualization of what it means to measure fidelity of implementation and a common method for doing so. Below I synthesize the arguments from these sources with respect to four ideas: 1) defining fidelity; 2) reasons for assessing fidelity; 3) criteria for assessing fidelity; and 4) methods for assessing fidelity.

Defining Fidelity

Although researchers have used a number of different terms (e.g., treatment integrity, compliance, adherence), the language of recent reviews suggests that "fidelity

of implementation" is now the most commonly applied label. Whereas the second most common phrase, "treatment integrity" (e.g., Schulte, Easton & Parker, 2009), suggests an inherent link to clinical trials, the phrase "fidelity of implementation" can be appropriately applied to instances of both research and practice. Although all of the reviewers reported that fidelity of implementation (FOI) has been defined in a multitude of ways, they converged on similar characterizations. Most broadly defined, O'Donnell (2008) employed the definition suggested by Loucks (1983): "the extent to which the user's current practice matche[s] the … ideal" (p. 4). Others operationalized "the ideal" as the "program model originally developed" (Mowbray et al. 2003). Dusenbury and colleagues' definition captures this theme in terms specific to educational interventions and serves as the definition I will employ in this proposal: "the degree to which teachers and other program providers implement programs *as intended by the program developers*" (p. 240, italics in original).

Reasons for Assessing Fidelity

Collectively, the five reviews provide ten reasons for assessing FOI. Although, in this dissertation, I will focus more on some than others, I consider all of the reasons to be valid and have synthesized them in an attempt to contribute to the building of a consensus regarding what it means to conduct fidelity studies.

The only reason about which all authors agreed is that of 1) *establishing causal relationships*. In the absence of data on the extent to which a program was implemented as intended, "researchers may not be able to account for negative or ambiguous findings,

Table 1 *Summary of reviews of fidelity of implementation (FOI)*

Source	Scope	Definition of FOI	Reasons for measuring FOI	Aspects of FOI to measure
Dane & Schneider (1998)	Reviewed 162 evaluations of primary and early secondary prevention programs (1980-1994)	(treatment integrity) "the degree to which specified procedures are implemented as planned" (p. 23)	o Establish causal relationships o Indications of feasibility o Gauge the effects of modifications o Program improvement	o Exposure o Adherence o Quality of delivery o Program differentiation o Participant responsiveness
Darrow (2009)	Reviewed 10 evaluation studies of 12 preschool curricula within the Preschool Curriculum Evaluation Research (PCER) Initiative	"Specific fidelity" (Hulleman & Cordray, 2009)	o Establish causal relationships o Indications of feasibility o Replication	o Exposure o Adherence o Participant responsiveness
Dusenbury et al. (2003)	Reviewed fidelity research in the fields of mental health, prevention of psychopathology, personal and social competence promotion, education, and drug abuse treatment and prevention for the purpose of providing fidelity guidelines for drug abuse prevention research.	"the degree to which teachers and other program providers implement programs *as intended by the program developers*" (p. 240, italics in original)	o Establish causal relationships o Indications of feasibility o Gauge the effects of modifications o Identify mechanisms of change	o Dose o Adherence o Quality of delivery o Program differentiation o Participant responsiveness
Mowbray et al. (2003)	Reviewed 21 studies (1987-2002) within the fields of mental health, substance abuse treatment, education, and social services.	"the extent to which delivery of an intervention adheres to the protocol or program model originally developed" (p. 315)	o Establish causal relationships o Replication o Improve power o Comparison of treatments o Guide for implementers	o Structure o Process
O'Donnell (2008)	Reviewed 23 studies in education (1977-2005) that "quantitatively measured the relationship between fidelity of implementation to K-12 core curriculum interventions and outcomes" (p. 37)	"'the extent to which the user's current practice matche[s] the … ideal' (Louchs, 1983, p. 4)"	o Establish causal relationships o Indications of feasibility o Improve power o Data exclusion	o Duration o Adherence o Quality of delivery o Program differentiation o Participant responsiveness

nor determine whether unsuccessful outcomes are due to an ineffective program or due to failure to implement the program and its conceptual and methodological underpinnings as intended" (O'Donnell, 2008, p. 42. In Bickman's (1987) terms, researchers would not be able to distinguish "program theory failure" from "program implementation failure." Furthermore, by collecting fidelity data, researchers may be able to identify those aspects of an intervention that have the greatest impact on the outcome(s) of interest (Darrow, 2009).

Four of the five reviews also suggested that studying FOI is important for *2) providing indications of feasibility* for implementing an intervention with fidelity in practice. For example, researchers can "identify components that appear to be important to success, but require additional support in order to be delivered in a high caliber fashion" (Darrow, 2009, p. 5). Implementation guidance that is differentiated by program components could inform practitioners' adoption decisions. Detailed descriptions of what is required to implement an intervention with fidelity could help potential implementers see that an intervention would not likely be implemented successfully in a particular setting (and so, they would avoid a questionable investment). Or, such descriptions could assist practitioners in allocating resources in a manner such that critical program components that are difficult to implement well would receive sufficient attention.

There was no consensus among the five reviews about other reasons for conducting FOI studies. Nonetheless, I include the eight additional reasons here as a background to my work. Reasons three through six pertain to methodological issues and have implications for the conduct of evaluation studies. The argument is made in two of the reviews that the design and results of FOI studies can support other researchers in *3)*

replicating an intervention in other settings because they necessarily provide detailed descriptions of the implementation process (Darrow, 2009; Mowbray et al., 2003). Additionally, when included as moderating variables in outcome studies, measures of fidelity can *4) improve statistical power* by helping to explain variance in outcomes (Mowbray et al., 2003; O'Donnell, 2008). Furthermore, measures of FOI can *5) provide a basis for excluding data* from settings in which the implementation of an intervention deviated too far from the original specification (O'Donnell, 2008) and *6) provide a means of comparing treatments*, whether they be variations of the same program or evaluations of multiple, similar programs (e.g., for conducting meta-analyses) (Mowbray et al., 2003).

The four remaining reasons for assessing FOI are relevant to the work of program developers and practitioners. Both Dane and Schneider (1998) and Dusenbury and colleagues (2003) argued that fidelity measures assist in *7) gauging the effects of modifications* to the prescribed activities of a program. Understanding both the effects of modifications and impediments to implementation can potentially contribute to *8) improving a program* (Dane & Schneider) and even *9) providing guidance for practitioners* in implementing a program as intended (Mowbray et al., 2003). Lastly, Dusenbury and colleagues suggested that because it "often helps to explain *why* innovations succeed and fail" (p. 240, italics in original), measuring FOI can assist in 10) *identifying a program's mechanisms for change.* Extending causal claims from explaining merely *that* A caused B to *why* A caused B affords opportunities for refining theory about how a program does or does not produce the desired outcome. But this is true only if, as argued by Hulleman and Cordray (2009), the evaluation begins with a

"well-stated set of expectations about how the intervention is supposed to work, its underlying logic, and rationales for how and why these actions will produce the desired enhancements in student learning, motivation, and achievement" (p. 90), which leads to a consideration of the FOI criteria that need to be measured in any fidelity study.

Criteria for Assessing Fidelity

Across the five reviews, the five criteria for assessing fidelity proposed by Dane and Schneider (1998) were consistently named as the important dimensions for which indices of FOI should be created: 1) exposure, 2) adherence, 3) quality of delivery, 4) program differentiation, and 5) participant responsiveness. However, the reviewers did not define these five criteria with the same consistency with which they named them as essential aspects of implementation. Across the various fields of interest, researchers' ways of labeling and conceptualizing these fidelity criteria vary. Therefore, because O'Donnell (2008) characterized FOI specifically for education research, her argument, which builds on those of Dane and Schneider and Mowbray and colleagues (2003), will be privileged here.

Based on their review of 162 evaluation studies of primary and early secondary prevention programs (e.g., substance-abuse prevention), Dane and Schneider (1998) named five aspects of FOI that should be included in authors' reports of examining the extent to which interventions were implemented as intended:

- *Exposure*—an index that may include any of the following: (a) the number of sessions implemented; (b) the length of each session; or (c) the frequency with which program techniques were implemented

- *Adherence*—the extent to which specified program components were delivered as prescribed in program manuals
- *Quality of delivery*—a measure of qualitative aspects of program delivery that are not directly related to the implementation of prescribed content, such as implementer enthusiasm, leader preparedness, global estimates of session effectiveness, and leader attitudes toward program
- *Program differentiation*—a manipulation check that is performed to safeguard against the diffusion of treatments, that is, to ensure that the subjects in each experimental condition received only planned interventions
- *Participant responsiveness*—a measure of participant response to program sessions, which may include indicators such as levels of participation and enthusiasm (p. 45)

O'Donnell's (2008) definitions of these five fidelity criteria are consistent with those of Dane and Schneider (1998), with one exception. Whereas Dane and Schneider's notion of quality of delivery—something "not directly related to the implementation of prescribed content"—seems more related to affect, O'Donnell defined this aspect of FOI as "the manner in which the implementer delivers the program using the techniques, processes, or methods prescribed" (p. 34), suggesting that the quality of delivery is potentially an integral component of the intervention. This subtle, yet significant, variation is likely a consequence of the domains within which the respective authors' conceived of FOI. Dane and Schneider were writing to a prevention sciences audience, O'Donnell to education researchers. As an illustration of these different conceptualizations of quality of delivery, I compare two studies reviewed by the authors. In the first, a case of a non-instructional intervention, Dane and Schneider view quality of delivery as potentially moderating the impact of FOI. In the second, where a particular form of pedagogy is fundamental to the program, O'Donnell uses quality of delivery to refer to whether the teacher delivered the units in the ways intended by developers.

Included in Dane and Schneider's (1998) review of prevention research was a report of a pilot-test of a school-based smoking prevention program conducted by Botvin,

Dusenbury, Baker, James-Ortiz, and Kerner (1989). The program targeted teenage students in urban schools and was implemented by regular classroom teachers. The teachers were provided with a one-day training workshop, during which they viewed a demonstration of and rehearsed the prevention curriculum activities. To measure the extent to which the teachers implemented the program as intended, trained observers collected two types of data from randomly selected class periods. First, the observers recorded which curriculum objectives were covered during a given session to later aggregate into a single index of completeness (i.e., proportion of curriculum covered). Second, using Likert-type scales, observers noted:

> the effectiveness with which the teacher implemented the program, the degree to which the teacher appeared prepared to implement the program, the attitude of the teacher toward the prevention program and toward students, and the extent to which the students actively participated in the program (p. 283).

The distinguishing feature of the program was that it approached the prevention of smoking from a psychosocial perspective (i.e., teaching social resistance skills) rather than more typical information dissemination strategies (Botvin et al., 1989). The skills the developers aimed to teach to students were sequenced in the program's curriculum. The greater the proportion of the curriculum that teachers covered, the more completely they implemented the intervention. Pedagogical guidelines for *how* the teachers were to teach the social resistance skills was not specified by program developers; rather, the qualitative aspects of teacher attitude and program use listed above were included by evaluators as potential moderators of program effectiveness. In that sense, Botvin and colleagues (as well as the reviewers, Dane and Schneider) reasoned that coverage of the

program's curriculum (i.e., perfect adherence to developers' specifications) depended on, for example, the teacher's level of excitement about the material or ability to relate to students.

In contrast, O'Donnell's (2008) review included a study conducted by Songer and Gotwals (2005) on the implementation of three inquiry-based science curricular units. Employing the National Research Council's (NRC, 1995) definition of scientific inquiry, the authors suggested that the intended manner of implementation of the units is modeled after scientists' methods of discovery. Rather than treating science as a "collection of irrefutable, disconnected facts" (p. 3), the curricula focused on "asking questions, exploring these questions, considering alternative explanations, and weighing evidence" (p. 3). To assess quality of delivery, the authors argued that the teachers' use of the curricular units would have to be directly observed to determine whether their instruction adhered to the forms called for by program developers.

Thus, whereas aspects of the quality of delivery of the smoking prevention program described above potentially moderate the impact of FOI, quality of delivery of the inquiry-based science units refers to whether the teacher delivered the units in the (qualitative) ways intended by developers, and is therefore *included in* (rather than moderating the impact of) FOI indices. This difference is consistent with a distinction that Mowbray and colleagues' (2003) made between fidelity to *structure*, the "framework for service delivery" (p. 318), and fidelity to *process*, which describes *how* services are to be delivered. O'Donnell argued that of the five aspects of FOI described above, exposure (which she relabeled 'duration') and adherence represent fidelity to structure, and quality of delivery and program differentiation fall into the category of fidelity to process (with

participant responsiveness taking on characteristics of fidelity to both structure and process). This structure-process distinction is helpful in identifying where the limits of program developers' prescriptions lie. For example, implementing the smoking prevention program described above in a manner faithful to developers' intentions required fidelity to only structure, whereas faithfully implementing the inquiry-based science curricular units required fidelity to both structure and process. Re-conceiving quality of delivery in these terms involves more than a shift in perspective. It represents a tailoring of FOI indices to programs that prescribe pedagogy, and has consequences for developing measures of FOI and interpreting results—a point to which I will return below when describing my instruments and methods in chapter three.

In the analyses of the relationship between FOI of Math Recovery and student outcomes reported below, I will use the language developed by the authors of the five reviews listed in table 1 above to describe the components of my fidelity coding scheme for the MR program, considering both structure- and process-oriented aspects of the MR model. First, however, I summarize the fourth and final idea concerning fidelity by outlining broad methodologically guidelines for conducting FOI assessments found in the literature.

Methods for Assessing Fidelity

Two of the reviews provided descriptions of concrete steps for conducting fidelity studies, the more detailed of which is that of O'Donnell (2008). Her six guidelines map roughly onto the five-step procedure proposed by Nelson, Cordray, Hulleman, Darrow,

and Sommer (2010) for assessing fidelity of implementation in evaluations of educational interventions. Figure 2 depicts the two schemes together. I follow each of O'Donnell's six guidelines here, discussing Nelson and colleagues' steps as they relate, supplemented with elaborations from other sources where appropriate.

O'Donnell (2008, p. 53)	**Nelson, Cordray, Hulleman, Darrow, & Sommer (2010)**
a) Establish the program theory a priori and determine what it means to implement the program with fidelity.	1) Specify the intervention model
b) Operationally define fidelity of implementation constructs and variables by specifying the critical components and process necessary for implementing the curriculum intervention with fidelity	
c) Develop separate instruments for measuring the critical components and processes. If the program promotes adaptation, measures of fidelity to the critical components and processes should be separate from measures of the user's adaptations and variations.	2) Identify appropriate fidelity indices
d) Incorporate random or full census sampling within the study in order to generalize fidelity findings to the study population.	
f) Test for and report on the reliability and validity of the fidelity data collected.	3) Determine index reliability and validity
	4) Combine indices where appropriate
e) Measure the user's fidelity to the critical components and processes; measure fidelity to processes in both the experimental and comparison condition, and relate these measures to outcomes.	5) Link fidelity to outcomes where possible

Figure 2. Guidelines/steps for assessing FOI proposed by O'Donnell (2008) and Nelson et al. (2010).

The first guideline follows previous arguments for the role of program theory in evaluations (Bickman, 1987; Lipsey, 1993), and in particular that FOI criteria and instruments should be based on the underlying theory of the program being evaluated.

That is, O'Donnell suggested that evaluators should begin by articulating a program's theory and "determine what it means to implement the program with fidelity" (p. 53). Cordray and Pion (1993) termed this the "rationale for change," and argued that "the form of the treatment intervention needs to be articulated in sufficient detail to be capable of specifying what is *unique* about the intervention (i.e., its active ingredients)" (p. 225, italics in original). This is what Fixsen and colleagues described as a program's "core components"—"the most essential and indispensable components of an intervention practice or program" (Fixsen, Naoom, Blasé, Friedman, & Wallace, 2005, p. 24). Such specification provides a basis from which to develop a FOI assessment scheme and instruments (Cordray & Pion, 1993). Furthermore, Connel and Kubisch (1999) suggested that results of evaluations designed with program theory in mind have implications for possible refinements to a program's "theory of change."

Second, program constructs and variables, including the necessary processes for implementing the program with fidelity, should be operationally defined (O'Donnell, 2008). Nelson and colleagues (2010) distinguished between an intervention's "change model" and its "logic model," with the former representing a network of causal connections between constructs (the focus of discussion in the preceding paragraph), and the latter consisting of "the resources and activities necessary to operationalize the change model components for the treatment condition of the experiment" (p. 15). However, in the case of unscripted interventions, some of which attempt to tailor instruction to individual student needs, translating the logic model into a plan for fidelity assessment is not always straightforward. Although an intervention itself might be adjusted to meet individuals' needs, an assessment of fidelity must be applied

consistently and systematically across all treatment cases. But Cordray and Pion (1993) argued that "[t]his form of tailoring does not mean that interventions are idiosyncratic. Rather, they usually involve an overarching [change] model (in the broadest sense) or treatment philosophy that dictates (or at least directs) the clinical course" (p. 230). With respect to identifying the resources and activities implicated in a model's clinical course, Waltz and colleagues (1993), from their work in assessing FOI in psychotherapy trials, suggested that implementers' behaviors fall into one of four categories with respect to a treatment program: (a) unique and essential; (b) essential but not unique; (c) compatible, but neither essential nor unique (and therefore not prohibited); or (d) prohibited (Waltz, Addis, Koerner, & Jacobson, 1993). Identifying behaviors that are (non)essential or (un)acceptable contributes to developing operational definitions for assessing FOI, particularly when evaluating unscripted interventions (Cordray & Pion, 2006).

The third step is that of developing instruments to document the implementation of core components and processes as defined in the previous step. O'Donnell (2008) argued that for programs in which adaptation is promoted, "measures of fidelity to the critical components and processes should be separate from measures of the user's adaptations and variations" (p. 53). Again, the categories identified by Waltz et al (1993) are potentially useful in making such distinctions, and in "indicating which components of the logic model are linked most closely with the 'core' components of the intervention, those that are essential to the theoretical process of the intervention that achieves its effects" (Nelson et al., p. 19, 2010).

Fourth, if FOI cannot be assessed for all participants, evaluators should randomly sample instances across the study so that findings with respect to fidelity can be

generalized to the entire study population. Fifth, evaluators should test and report the reliability and validity of their instruments and the fidelity data collected.

The sixth guideline pertains to data collection and analysis. After assessing all five fidelity criteria defined above, each should be related to outcomes where possible. As Nelson et al. (2010) pointed out, an intervention implemented with consistently high fidelity will have insufficient variation in fidelity indices for examining relationships between FOI and outcomes. But, such implementation success in field experiments is rare. There is typically plenty of variation in FOI to measure, report, and use in analyses. O'Donnell (2008) argued that too often researchers report 'monitoring' *structural* aspects of fidelity without assessing users' fidelity to program *processes* and, in so doing, fail to account for the variation in FOI that is most strongly related to outcomes (Mowbray et al., 2003). Additionally, O'Donnell argued that fidelity and adaptation should be treated as separate constructs in analyses, with their relationships to outcomes analyzed separately.

Taken together, O'Donnell's (2008) and Nelson et al.'s (2010) guidelines form a general action plan for conducting fidelity studies in education evaluation research, applicable to a wide range of interventions that differ in terms of the extent to which they are scripted, adaptability, scale and duration.

Summary

In this chapter I have synthesized five recent reviews of FOI work with respect to four major issues: 1) defining fidelity; 2) reasons for assessing fidelity; 3) criteria for

assessing fidelity; and 4) methods for assessing fidelity. I interpreted the findings of these reviews as evidence that a consensus is emerging about role of fidelity of implementation in evaluation research. This consensus is aligning with a program theory perspective such as that articulated by Cordray & Pion (1993):

> The overall assessment and research design must map onto the structure and operations of the program throughout all its stages… Although not always well articulated, interventions do not "spring out of the blue"; rather, they are grounded in some notion about why the services to be delivered should remedy the targeted problem(s). To determine whether the mechanisms underlying the stated rationale are initiated by the intervention, assessment is needed (p. 224-225).

I argue that not only does such a perspective extend to assessing fidelity, but that fidelity studies are necessary if evaluation studies are to examine and test theorized change mechanisms. However, while there are indications that a consensus is emerging with respect to the first three issues listed above, the fourth, methods for assessing FOI, has not received such attention. There does not yet exist a standard set of methods for conducting fidelity studies and linking their results to study outcomes.

This dissertation is to address this limitation, particularly with respect to unscripted interventions. The four purposes of this dissertation are to 1) assess the extent to which Math Recovery was implemented with fidelity during the evaluation study in order to determine whether it was/can be implemented with sufficiently high fidelity; 2) link fidelity indices to student outcomes in order to test the soundness of MR's program theory; 3) use the results to provide guidance to program developers in improving the intervention; and 4) articulate methods for assessing FOI of a particular type of

intervention—one that is highly unscripted and in which adaptation is encouraged. I will address all of these purposes by applying the framework developed in the above section. Specifically, I trace my work through O'Donnell's (2008) six guidelines, describing my methods for assessing fidelity to both structural and process components of MR (Mowbray et al., 2003), including the fidelity criteria identified by Dane and Schneider (1998). However, before doing so, I describe the MR program and the larger evaluation project within which the fidelity study was conducted.

CHAPTER 3

METHODS

In this chapter, I document my methods for conducting the fidelity study of the implementation of Math Recovery and for examining the relationship between FOI and student outcomes. As stated at the end of Chapter 2 my overarching goals include: 1) assessing MR's potential for successful scale-up (i.e., determining whether it was successfully implemented with fidelity), 2) testing MR's underlying program theory, and 3) identifying potential areas for program improvement. Additionally, I intend to 4) articulate my methods for assessing the FOI of a particular type of intervention—one that is highly unscripted and in which adaptation is encouraged—in order to provide guidance to others doing similar work.

In the first section, I provide a background to the fidelity study. I first describe the MR intervention in some detail. Providing a full explication of the theory underlying the MR program is consistent with the program theory perspective I outlined in Chapter 2 (Bickman, 1987; Connel & Kubisch, 1999; Cordray & Pion, 1993; Lipsey, 1993; O'Donnell, 2008). I will draw on this explication when justifying the aspects of MR that I identified as important to measure, the instruments I developed for doing so, and the models with which I assessed the relationship between FOI and student outcomes. Following a description of MR, I provide an account of the larger evaluation study within which I conducted the fidelity study.

In the next section of the chapter, I step through the first five (of six) guidelines for assessing FOI provided by O'Donnell (2008) and Nelson and colleagues (2010), which I outlined in Chapter 2. These include: 1) identifying the program theory and core components; 2) operationally defining program constructs and variables; 3) developing instruments; 4) sampling; and 5) determining and reporting instrument reliability validity. Additionally, I describe the hiring and training of coders.

In the final section of the chapter, I address O'Donnell's (2008) and Nelson et al.'s (2010) sixth guideline by describing my analyses. In doing so, I address three sets of issues: (a) the creation and properties of the FOI indices used in the analyses, including variable construction, distribution and variance composition; (b) rater effect; and (c) the research questions, including descriptions of the models employed to answer those questions.

Background to the Fidelity Study

Description of Math Recovery

Math Recovery is a diagnostic, pullout tutoring intervention aimed at increasing the mathematics achievement of low-performing first graders. The goal of MR is to close the persistent pre-K achievement gap in mathematics (Aubrey, Dahl, & Godfrey, 2006; Duncan, Claessens, & Engel, 2004; Princiotta, Flanagan, & Germino Hausken, 2006). By providing near-daily, one-to-one instruction for approximately one third of the school year, developers of the MR intervention aim to enable the lowest achieving first-graders to achieve in the regular mathematics classroom at levels comparable to those of their

higher-performing peers, and to retain these gains in subsequent grades (with no need for follow-up tutoring). Determination of eligibility is based on teacher recommendation and an initial screening process. Generally, the population targeted by the program is students performing in the bottom quartile of first graders in mathematics in a school who are not supported by special education services.

Broadly speaking, MR consists of three primary components: 1) tutor training, 2) student identification and assessment, and 3) one-to-one tutoring. After briefly describing the origins of MR, I discuss the second and third components in detail, including an illustration of what MR tutors are expected to do. Then, I describe the training that MR provides tutors to support them in developing such practices.

The MR program is based on research and a developmental framework that was developed to study the *process* of young children's arithmetical learning (Steffe, von Glasersfeld, Richards, & Cobb, 1983; Steffe, Cobb, & von Glasersfeld, 1988). This one-to-one teaching methodology was itself developed as an adaptation of Piagetian clinical interviewing. The interviewer's intent in conducting a clinical interview is to assess the development level of the child's arithmetical reasoning. The interviewer ensures that the child has every opportunity to understand the intent of the tasks posed *but does not cue* the child to the correct response or otherwise intervene to support the child's solution process. The clinical interviewing methodology enables researchers to construct a sequence of "snap shots" in the development of children's reasoning but it does not allow them to study the process by which children make the transition from a less sophisticated to a more sophisticated level. The one-to-one teaching experiment methodology was designed to overcome this limitation. In conducting an investigation of this type, the

researcher teaches a small number of children one-to-one on a regular basis for an extended period of time. Throughout the sessions, the researcher continually assesses each child's mathematical reasoning and uses these assessments to inform the selection of tasks designed to support the child's learning. The researcher's goal is to pose tasks that are cognitively demanding for the child but that the child can engage in meaningfully (i.e., not too easy or too difficult at this particular point in the child's development). In addition to posing developmentally appropriate tasks, the primary means the researcher uses to support the child's learning is to encourage the child to reflect on and modify his or her ongoing problem solving activity. The key point to emphasize is that the immediate objective of a one-to-one teaching experiment is to develop cognitive models of the process of children's arithmetical learning rather than to improve either instructional task sequences or teaching practice.

The role of tutors in the MR program is similar to that of a researcher conducting a one-to-one teaching experiment. The tutor works with children individually and conducts initial and ongoing assessments that guide the selection of tasks and the posing of follow-up questions designed to support the children's reflection on their problem-solving activity. This approach to teaching reflects two core assumptions: significant mathematical learning occurs only when 1) children encounter a situation in which their current concepts and solutions methods prove to be inadequate, and 2) children can resolve these difficulties with only limited support from a teacher or more knowledgeable other. For example, during MR tutors training, trainers emphasize that in tasks involving problem solving, MR tutors are not to tell students whether their solutions are correct or incorrect, but should instead continually press students to explain their thinking and

require students to check their own answers. In employing this assessment approach, tutors are to refrain from acting as the mathematical authority in interactions with students. Rather, in keeping with the core assumptions of the program, and in order to support students in developing confidence in their own mathematical capabilities, the goal of MR is to provide students with opportunities to think for extended periods of time about problems and come to their own conclusions and to develop ways of validating (or invalidating) those conclusions.

Operationally, there are two *unique aspects* of MR tutoring as compared to typical mathematics interventions. These entail: 1) a tutor's *ongoing assessment* of the child's thinking and strategies (both retrospective assessment between tutoring sessions and in-the-moment assessment during tutoring sessions), and 2) a tutor's efforts to pose tasks at the high-end margin of the child's *zone of proximal development*.

Much of MR's underlying theory for how tutors accomplish the goal of assessing children's arithmetical knowledge and tailoring instruction to meet their current needs is drawn from cognitive research on children's development of arithmetical reasoning. The developers of MR have codified this knowledge in two frameworks: the MR Learning Framework in Number and the MR Instructional Framework in Early Number (Phillips, Leonard, Horton, Wright, & Stafford, 2003; Wright, 2003; Wright, Martland, & Stafford, 2006; Wright, Martland, Stafford, & Stanger, 2006). The MR Learning Framework is based on cognitive models of children's numerical reasoning proposed by a number of researchers (Baroody, 1987; Baroody & Ginsburg, 1986; Carpenter & Moser, 1982, 1984; Fuson, 1988, 1992; Steffe, Cobb, & von Glasersfeld, 1988; Steffe, von Glasersfeld, Richards, & Cobb, 1983). The Learning Framework is designed to assist the tutor in

diagnosing students' current understandings and strategies by providing information about students' responses in terms of levels of sophistication. The tutor uses this framework to create (and update daily) a profile for each student. The profile includes ratings of a student's current arithmetical strategies and knowledge of early number for each of six aspects of early number. These aspects are defined in Figure 3, along with descriptions of their respective ranges of stages/levels.

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

Figure 3. Aspects of early number knowledge included in MR Learning Framework

The Instructional Framework is based on research on early number instruction (Baroody, 1990; Beishuizen, 1993; Carpenter, Franke, Jacobs, Fennema, & Empson, 1997; Cobb, Gravemeijer, Yackel, McClain, & Whitenack, 1997; Fuson, 1990; Fuson, Wearne, Hiebert, Human, et al., 1997; Hiebert & Wearne, 1992). The tutor uses the MR Instructional Framework to develop an instructional plan that is based on the child's profile. The Instructional Framework specifies at least six exemplary "procedures" for each of 30 key topics in arithmetic that map onto the levels of the six aspects of early number knowledge specified in the Learning Framework. Each procedure, as described in the MR instructional handbook (Wright, Martland, Stafford, & Stanger, 2006), consists of a set of tasks, accompanied by a description of the appropriate tools or manipulatives to use when posing the tasks. Additionally, each procedure provides detailed guidance for the tutor and includes a statement of its purpose, specification of the teacher's words and actions, a detailed discussion of possible student responses, and descriptions of applicable instructional materials. It is important to note, however, that the tutor is not limited to the procedures described in the handbook; tutors may employ tasks from other sources as long as those tasks align with the Instructional Framework and match the intent and spirit of those described in the MR handbook. As described by the authors, the procedures are "intended to be illustrative and are not intended necessarily to be followed verbatim" (p. 71).

As an illustration of the work tutors must do to use the MR frameworks to diagnose students' strategies and pose appropriate tasks, consider the following scenario, which pertains to a hypothetical student's strategy for determining the sum of two addends (included in the SEAL aspect of the student's profile described in Figure 3).

Based on the initial assessment, which includes 17 sections, (or tasks posed during a tutoring session), a tutor discovers that when asked to find the sum of nine red chips and four blue chips placed on a table with each set covered by a screen, the student can successfully solve the task using a "count from one" strategy. That is, the student first counts the hidden red chips from one to nine and then continues the count from ten to thirteen to include the hidden blue chips as well. The tutor would need to locate this strategy on the MR Learning Framework as representing "Stage 2," and then link this aspect of the student's profile to the Instructional Framework to determine the kinds of tasks (s)he should pose to support the student in developing a more sophisticated ("Stage 3") strategy. In this case, the goal would be to support the student in developing a "counting on" strategy. Instead of counting each of the nine red chips, the student at Stage 3 would simply begin with the first quantity and count the second ("nine… ten, eleven, twelve, thirteen!"). To support the student in developing such counting-on strategy, the tutor would be directed to employ a similar task (with two screened quantities), but to increase the size of the first addend (e.g., 23 + 4, or 41 + 2). The rationale is that, as the tutor poses tasks with larger and larger first addends but consistently small second addends, the student will cease to employ the increasingly time-consuming step of counting each chip of in the first set, and realize (s)he can just count on.

According to the MR model, each child selected for MR tutoring receives 4-5 one-to-one tutoring sessions of 25-30 minutes each week for approximately 11 weeks. Every tutoring sessions is video-recorded for two purposes: 1) the tutor reviews each recording before the next day's lesson to update the student's profile and plan the

following session by selecting appropriate tasks; and 2) because each tutoring session can later be observed on video, the tutor can remain 'in the moment' throughout a session, focusing exclusively on continually assessing the student's thinking and responding by posing appropriate tasks. Each lesson consists of a variety of procedures representing several (though usually not all) of the six aspects of early number listed in Figure 3.

In most schools and districts that implement MR, teachers are employed as half-day MR tutors, working with only 3-4 students at any given time to ensure that they have enough time to review video recordings and plan for subsequent lessons. Generally, there are three tutoring cycles per school year (with 3-4 students per cycle), so a MR tutor will typically serve 9-12 students per year.

Implementing MR first requires selecting and training tutors. A primary goal of tutor training is supporting MR tutors in comprehending and effectively using the three fundamental tools with which they are equipped: 1) the MR Learning Framework; 2) the MR Instructional Framework; and 3) the MR handbook of recommended, exemplary teaching tasks mapped onto the Instructional Framework. To support MR tutors in using these tools effectively, tutors receive 60 hours of training conducted by a MR leader, which addresses the theory and techniques of the MR program. The overall goal of MR training is to enable teachers to use a range of pedagogical tools for clinical assessment and intervention that include MR assessments, the Learning Framework, and the Instructional Framework. The training typically involves an initial five-day workshop during the summer and a follow-up three-day workshop conducted three weeks into the school year after trainees have completed the initial MR assessments of students selected for the program. During training sessions, teachers view excerpts of video-recorded

assessment interviews and tutoring sessions selected for training purposes. The MR leader guides the trainees' interpretation of these excerpts by orienting them to focus on critical aspects of the tutor's and child's behaviors. Additionally, the MR leader teaches trainees to review video recordings of assessment and tutoring sessions in order to improve their own assessment and tutoring practices. The trainees in a district are also expected to meet as a cohort for two hours each month during the school year. The MR leader responsible for their training typically attends three of these meetings and also conducts individual coaching sessions with the MR trainees during these site visits.

Evaluating Math Recovery

The two-year, randomized field trial of Math Recovery (of which the fidelity study reported here was a part) was conducted in 20 elementary schools (five urban, ten suburban and five rural), representing five districts in two states. Each was a 'fresh site' in that the program was implemented for the first time for the purposes of the study. Eighteen teachers were recruited to receive training and participate as MR tutors from the participating districts. All tutors had at least two years of elementary classroom teaching experience. Sixteen of the tutors received half-time teaching releases to serve one school each; two of the tutors served two schools each, tutoring all day. Each participating school district received $5000 per tutor per year for the two years of the study and free training for tutors, with remaining costs underwritten by the districts. The same MR trainers provided training in each of the two states; five teachers from the five rural districts were trained in one site, and thirteen teachers from the urban and suburban districts were trained in another.

In each of the two years of the evaluation study (2007-08 and 2008-09 academic years), 17 to 36 first graders in each school were identified as eligible at the start of first grade based on their performance on the initial MR assessment. Eligible students in each school were randomly assigned to one of three tutoring cohorts corresponding to the typical three MR tutoring cycles per year, or to a "wait list" for MR. The cohorts, consisting of three students each, were staggered across the school year (i.e., Cohort A—September, B—December, C—March). In both years of the study, students on the randomly ordered waiting list were selected to join an MR tutoring cohort if an assigned participant left the school or was deemed ineligible due to a special education placement. The number of study participants before attrition totaled 517 in Year 1 and 510 in Year 2, of whom 172 received tutoring in Year 1 and 171 received tutoring in Year 2. Additionally, we followed students who participated in Year 1 through the second year of the study to administer one additonal assessment at the end of second grade. Approximately 50% of participants were males, 48% were non-white, and 48% received free or reduced lunch.

As noted above, the MR program is structured so that three tutoring cycles are conducted each school year. This allowed us to use the fact that, for two thirds of the participating students, treatment was delayed by either 11 or 22 weeks to establish an experimentally assigned control group for each cohort of participants, consisting of both students whose treatment had not yet begun and the students on the "wait list" for treatment. By randomly assigning the students selected for participation in the study each year to one of the three treatment cohorts or the wait list, the essential characteristics of an experimental design were established: a comparison of students' change in

mathematics achievement during their 12 weeks of participation in MR to the gains they would have made had they not received tutoring.

The three waves of treatment each year required four points of assessment—a pretest at the beginning of the school year and posttests immediately following each of the three cycles of tutoring. Of course, not all students assigned to treatment were ending or about to begin tutoring. Therefore, either a full battery or a partial battery of assessments was administered to students depending on whether they were entering or exiting tutoring at a particular time point. The full battery of measures was administered to all participating students at the beginning and end of the first grade year and, for those students who were participants in the first year, at the end of the second grade year. Additionally, students assigned to treatment cohorts completed the full battery of assessments when entering a tutoring cycle (which coincided with the beginning-of-year pretest for cohort 1 of each year) or exiting a tutoring cycle (which coincided with the end-of-year posttest for cohort 3 of each year). The partial battery was administered only at time points two and three each year—to all waitlist students and to students who were not entering or exiting a tutoring cycle.

The full battery consisted of five measures of mathematics knowledge, including three off-the-shelf, nationally normed assessments, an assessment that was part of the MR program, and an assessment that was developed specifically for the evaluation study. The three nationally normed assessments were alternating forms of the Applied Problems, Quantitative Concepts, and Fluency subtests of the Woodcock-Johnson III Achievement Test (hereafter referred to as WJ III). These assessments have been demonstrated to be sufficiently valid and reliable. The median reliability coefficient alphas for all age groups

for the standard battery of the WJ III for tests ranged from .81 to .94. Test items were developed with contributions from experts and were designed to measure both narrow and broad abilities. Concurrent validity of the WJ III was demonstrated 1) by showing that tests from the same cluster are highly and significantly correlated and those from different clusters correlate at a lower level and 2) by demonstrating correlations with other validated assessments (Woodcock, McGrew & Mather, 2001). The assessment that was part of the program was the initial MR assessment administered by tutors and described above. The last assessment, the 'MR Proximal,' was an instrument developed in consultation with the program developers. First and second grade versions of the MR Proximal were designed specifically to measure what students would likely learn in the course of the MR intervention in the case that the WJ III lacked sufficient sensitivity. Three parallel forms of the MR Proximal were used. The three Woodcock Johnson subtests and the MR Proximal instrument were administered by external assessors hired by the research team whereas the tutors administered the initial MR assessment.

The partial battery of assessments included only the WJ III Fluency subtest and the MR Proximal. With the exception of the initial MR assessment, which was always administered by tutors, all of the measures were administered by external assessors—retired teachers, hired and trained by the evaluation team.

Assessing Fidelity of Implementation of Math Recovery

Having provided a description of the MR intervention and the larger evaluation within which the fidelity assessment was conducted, I now trace my methods for

assessing FOI and its impact on evaluation outcomes through O'Donnell's (2008) and Nelson et al.'s (2010) six guidelines defined in Chapter 2: 1) identifying the program theory and core components; 2) operationally defining program constructs and variables; 3) developing instruments; 4) sampling; 5) determining and reporting instrument reliability validity; and 6) analysis. I also describe the hiring and training of coders. My goal in doing so is to explain how I assessed fidelity to both structural and process components of MR (Mowbray et al., 2003), including the fidelity criteria identified by Dane and Schneider (1998).

Math Recovery Program Theory and Core Components

As stated above, the goal of the MR program is to enable the lowest achieving first-graders to achieve in the regular mathematics classroom at levels comparable to those of their higher-performing peers, and to continue to do so in subsequent grades. Inherent in the MR design is the notion that students who are low-performing in mathematics have not had the opportunity to think deeply about number or to develop effective strategies that make sense to them. For example, in the regular mathematics classroom, their higher-performing peers might consistently voice solutions to problems before students eligible for MR have had time to think through the problem completely. According to MR developers, providing intensive, tailored instruction over the course of multiple weeks will provide such opportunities efficiently because such instruction is based on continuous assessments of the student's current understanding and is intended to always be at the "cutting edge" of that understanding. That is, little time in MR tutoring is spent practicing ideas and procedures that students have previously mastered. According

to the MR model, tasks should be limited to those that are within a student's zone of proximal development (ZPD). In MR language, they should be "genuine problems" for students—tasks that require students to think in ways that are just beyond the strategies they have previously demonstrated. The MR model is premised on the assumption that students can learn while solving such tasks with very limited scaffolding by the tutor (i.e., the selection and posing of tasks, and the provision of manipulatives when judged appropriate).

Thus, the means by which Math Recovery is theorized to "work" can be succinctly stated as the following. After training, MR tutors should be able to effectively employ research-based models of elementary students' learning progressions in early number (i.e., the MR Learning Framework) and means of supporting those progressions instructionally (i.e., the MR Instructional Framework) to diagnose students' current understanding and pose appropriate tasks that are in the student's zone of proximal development. This requires tutors to make an accurate assessment of the student's knowledge upon entering a tutoring cycle and to continually update their assessment both during and after tutoring sessions. MR developers contend that, as a result of the intervention, students will develop fluency in identifying symbols and number word sequences, and develop arithmetical strategies that will both enable them to participate successfully in the regular mathematics classroom after tutoring has ended and provide a foundation for learning in subsequent grades. Figure 4 depicts this change model.

Having articulated MR's program theory, the next step in following O'Donnell (2008) and Nelson and colleagues' (2010) guidelines above was to identify its core components. Before the evaluation began, I examined MR materials, including published

handbooks (Wright, Martland, & Stafford, 2006; Wright, Martland, Stafford, & Stanger, 2006) and resources provided by the U.S. Math Recovery Council (USMRC) (e.g., training manuals, instructional manipulatives, etc.), attended regional MR conferences, and consulted program developers—all in an attempt to identify the core components of

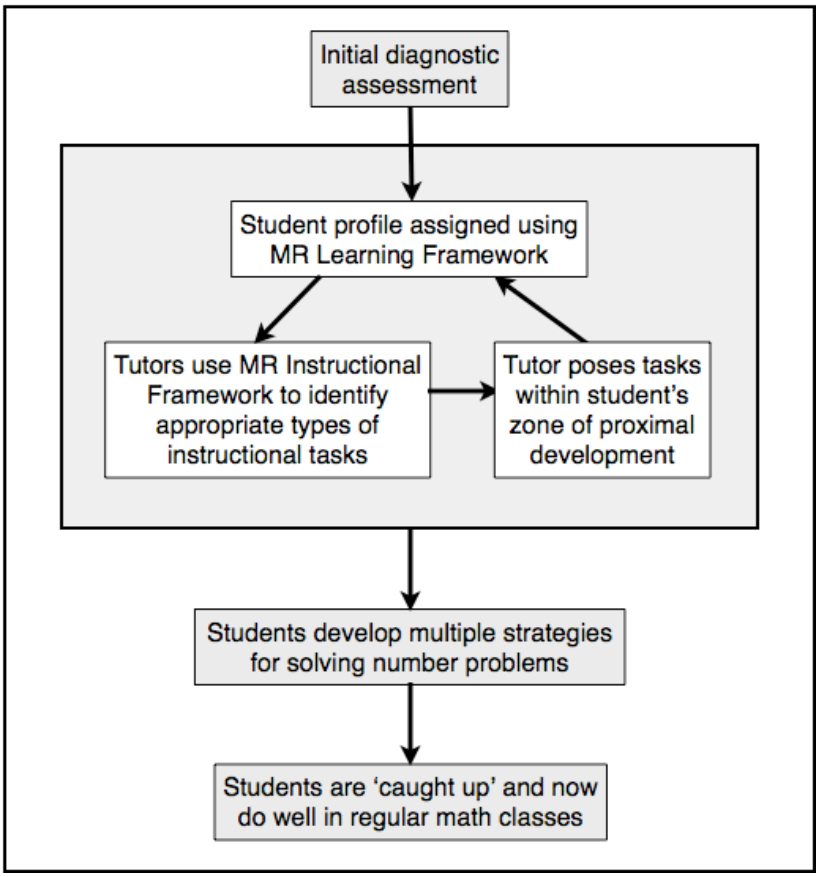


Figure 4. MR Change Model.

Math Recovery and develop initial schemes for assessing those constructs. Guiding the fidelity assessment in the end were what we, in collaboration with program developers, determined to be "unique and essential" and "essential but not unique" elements of MR, as well as "prohibited" behaviors (Waltz et al., 1993).

Essential components of the program that are unique to MR tutoring (as compared to typical tutoring) include: (a) the tutor's ongoing assessment of the student's thinking and strategies (both reflective assessment between tutoring sessions and in-the-moment assessment); and (b) the tutor's efforts to provide tasks within the student's zone of proximal development. Essential components that are not unique to MR include: (a) providing the student with sufficient time to think about and solve a task (i.e., "post-task wait time"); (b) requiring the student to check his/her own solutions by employing a different strategy than the one employed to reach the original solution ("child checking"); and (c) soliciting an explanation from the student after he or she has solved a task, to provide the tutor with evidence of the student's strategy and to support the student in reflecting on the solution ("solicitation of student strategy"). Finally, prohibited tutor behaviors include: (a) directly demonstrating a method for solving a problem (rather than allowing the student to develop and use his/her own strategies); and (b) eliciting behaviors from the student (e.g., entering into guessing games in which the student attempts to be effective by figuring out what the tutor has in mind rather than by thinking through a problem). In the next section, I describe my operational definitions of these core components.

Operational Definitions of MR Constructs and Indicators

Operational definitions for each of the core components listed above are presented in Figure 5. The components are organized by whether they pertain to the initial assessment or to instructional sessions. Additionally, the components' measures are organized by fidelity criteria (i.e., exposure, adherence, participant responsiveness,

quality of delivery, and program differentiation), which are grouped within structure- and process-oriented categories. I discuss each of these categorizations as I define the components to show how the fidelity indices account for aspects of FOI identified in the literature. For example, I have categorized most of the elements that pertain to the initial assessment as "adherence" aspects of fidelity. This is because the initial assessment is much more scripted than the tutoring sessions. The program developers expect tutors to follow a particular sequence of tasks, grouped in 17 sections, to generate information about the student's current understanding of each aspect of the MR Learning Framework that will enable them to assign a profile on which initial instructional choices will then be based. Therefore, administering the initial MR assessment is predominantly about following a set of directions.

The operationalizations described in this section cover all of the FOI indices that I created. As I will clarify, it was not possible to code all the indices with sufficient reliability and, therefore, they were not all included in the final models for linking FOI indices to student outcomes. I nevertheless provide an exhaustive description here in order to document the full scope of my work and to highlight those areas that proved to be particularly challenging.

Regarding tutors' administration of the initial assessment, the primary components that it was important to consider included (a) whether the assessment was indeed administered; (b) the frequency with which tutors committed errors in administering the assessment (defined below); (c) whether the tutor administered tasks and posed follow-up probes to produce a sufficient amount of information to document the limits of the child's current thinking in each of the aspects of the MR Learning

Core MR Components		Measures	CATEGORY	Structure		Process		
			FOI ASPECT: Exposure/duration; **A**dherence; **P**articipant **R**esponsiveness; **Q**uality of delivery; **P**rogram **D**ifferentiation →	**E**	**A**	**PR**	**Q**	**PD**
			Definition					
Initial Assessment		Use of initial assessment	Whether the 1.1 and 2.1 assessment were given at the outset of a tutoring cycle.		•			
		Error rate	Average frequency of error on portions of the assessment, including both 'major' and 'minor' errors.		•		•	
		Sufficient information revealed	Number of five strands of the MR Learning Framework on which the initial assessments generate enough information from the students' response to create a profile for that student's current thinking			•		
		Correct Profile/ Portrait	Whether external assessment agrees with tutor's profile/portrait of student		•			
Instruction		Frequency/duration of instruction	(a) the total number of tutoring sessions provided (exposure); (b) the average length of each session (duration); and (c) the average (per lesson) amount of time spent on strategy-based teaching procedures	•				
	Ongoing Assessment	Adjustment of task challenge	After tutor poses task & student cannot answer correctly (particularly cases where the student is not using any strategy), whether tutor reduces level of difficulty of task (within a task)				•	
		Responsiveness to student thinking/ answer	Whether tutor's subsequent instructional choices are in response to student's strategies/thinking on previous strategy-focused task. (If two tasks are posed together, rate only the first task.)				•	
		Evidence of student strategy	Whether, based on observation of student solving task and/or response to tutor's explicit solicitation, student thinking/strategy was evident			•		
		Correct Profile/ Portrait	Whether external assessment agrees with tutor's profile/portrait of student		•			
	ZPD	Task challenge	Extent to which task afforded student opportunity to struggle meaningfully with mathematics; it was a "genuine problem." This rates what the task became if there was adjustment.				•	
		Teaching procedure-portrait alignment	Whether teaching procedure targets the progression from one Learning Framework stage/level to the next		•			
	Nature of Instruction	Post-task wait-time	Whether tutor allows sufficient time for student to think/problem solve/answer				•	
		Child Checking	When tutor asks/allows student to check his/her (last) answer (typically with a reduction in difficulty), whether this is after a correct response or an incorrect response.				•	
		Solicitation of student strategy	When tutor explicitly asks student to explain strategy/thinking, whether this is after a correct response or an incorrect response.				•	
		Method demonstration	Whether tutor (un)intentionally demonstrates the method for how to solve a task				•	
		Behavior eliciting	Whether tutor directs the student to solve a task (or at least a step) in a particular way; or emphasizes a particular solution method or response				•	
	Positive Infidelity	Re-voicing	Whether tutor re-voices student's strategy				•	
		Different strategy	Whether tutor asks student to solve problem in a different way				•	
		Compare strategies	Whether tutor's questions encourage student to examine the mathematical similarities and differences among two or more strategies				•	

Figure 5. Operational definitions of core components of MR instruction.

Framework (i.e., the number of aspects of the MR Learning Framework on which a student's profile could be determined based on the tutor's use of the initial assessment); and (d) whether the tutor assigned the correct student profile based on the initial assessment. Regarding the second of these four components of the initial assessment, I defined what constituted a *minor error*, a *major error*, or *no error* in consultation with MR experts[1]. For example, minor errors included instances in which a tutor changed a particular quantity indicated in a task on the assessment protocol without changing the spirit of assessment (e.g., asked the student to add 8 and 3 more, instead of 8 and 4 more). Major errors included instances when tutors used the introductory task (which is intended as simply a 'warm-up' into that portion of the assessment) to "teach" a method for solving such tasks. However, the MR experts indicated that giving a student the answer to an introductory task after he or she had attempted to solve it was not classified as an error.

As indicated in Figure 5, because errors could potentially represent a failure to adhere to the protocol (e.g., changing or skipping tasks), or a violation of the intended delivery (e.g., "teaching" rather than merely assessing), this component pertains to two fidelity criteria: adherence and quality of delivery. Similarly, the third component, whether the tutor produces sufficient information about the student's thinking during a tutoring session, also requires attention to the student's contributions.

Regarding tutoring or instructional sessions, I categorized the core components into five subgroups. The first includes only measures of exposure and duration: the total number of tutoring sessions conducted with a student (exposure); the average length of

[1] This included three MR experts: the lead developer of the program and two of its primary trainers in the U.S. (who, as explained below, provided the training to our coders).

each session (duration); and the average (per session) amount of time spent on tasks designed to support strategy development (the rationale for this measure is explained below).

The second and third subgroups both specify the two unique and essential components listed above: conducting ongoing assessments of the student's knowledge, and the posing of tasks that are within a student's ZPD. The assessment subgroup includes four aspects: (a) whether the tutor assigned the correct student profile based on the results of the previous lesson(s); (b) whether, based on observation of the student solving a task and/or response to the tutor's explicit solicitation, there was evidence of the strategy the student employed to solve a task; (c) whether the instructional choices the tutor subsequently made were in response to the student's strategies/thinking on the previous task; and (d) whether, in instances when a task was too difficult for a student, the tutor responded by adjusting that task's level of difficulty. The first two aspects parallel the corresponding aspects of the initial assessment described above. The first aspect concerns whether the tutor correctly used the MR Learning Framework to assign a profile for the student, and is therefore an issue of adherence. (Unfortunately, however, tutors' paper records, on which they are expected to update students' Learning Framework profiles daily and create an appropriate lesson plan, were not kept with sufficient consistency to address this first aspect. Therefore, we could not assess the accuracy or diligence with which tutors updated students' profiles using the MR Learning Framework.) The second aspect, whether there was evidence of the strategy the student employed to solve a task, requires attention to the student's contribution, and therefore pertains to student responsiveness. Because the third and fourth aspects pertain to *how* the

tutor uses the results of previous tasks to inform subsequent task selection, or handles instances of the student reaching an impasse, they pertain to quality of delivery.

The third subgroup, which focuses on whether the tutored posed tasks that are within a student's ZPD, includes two aspects: (a) the extent to which a task afforded the student an opportunity to engage meaningfully with mathematics (i.e., the task was a "genuine problem"); and (b) whether a set of tasks targeted the progression from one Learning Framework stage/level to the next. Because the first of these two aspects requires the coder to make a judgment of the *quality* of each task the tutor chooses to provide a student (a choice in which the tutor is given considerable leeway in the MR program), I view it as pertaining to quality of delivery. The second of the aspects concerns whether the tutor used the MR Instructional Framework correctly and therefore pertains to adherence.

The fourth subgroup of aspects of MR instruction, nature of instruction, includes all of the essential but not unique aspects listed above ("post-task wait time," "child checking," and "solicitation of student strategy") as well as behaviors explicitly prohibited by the MR model (directly demonstrating a method for solving a problem and behavior-eliciting). In contrast to the other aspects in this category, child checking (requiring the student to check his/her own solutions by employing a different strategy than that used to solve the task initially) and solicitation of student strategy (soliciting an explanation from the student after solving a task) are defined not solely by *whether* they occurred, but also *when*. In other words, the issue is not the frequency with which a tutor asks a student to check her/his work or explicitly asks the student to explain the strategy (s)he has used to solve a task. Instead, it is whether there is a pattern to the tutor's uses of

these moves: if the tutor asks a student to check her/his work after only incorrect responses, then the student might learn to recognize such a prompt as an indication that her or his response is incorrect and that (s)he should produce (or guess) a different answer. In this case, the student's learning might not be solely about developing new arithmetical strategies, but could be primarily about figuring out what the tutor has in mind.

Taken together, the components in the fourth subgroup provide an excellent illustration of the re-conceptualization of the "quality of delivery" aspect of FOI for evaluating educational programs that prescribe pedagogy. *How* the tutors are to deliver the program is fundamental to the program, not merely a potential moderating variable. For example, if a tutor consistently demonstrates methods for students to use, (s)he is not only potentially diminishing or amplifying the effects of MR, but is acting in direct contradiction to the program developers' expectations. In could be argued that this is simply a matter of adherence—in violating a rule of the MR program, the tutor is not adhering to the model. I would argue that how these aspects are labeled is not as important as *how they are assessed.* One reason for including them in the "quality of delivery" aspect of FOI is to emphasize their importance with respect to other fidelity indices. Making reliable qualitative judgments about the nature of tutors' instruction is not an easy task—one that evaluators frequently avoid. But, as mentioned above, such aspects represent the very kind of *process* variables that likely have strong relationships to outcomes, but, as O'Donnell (2008) argued, are too often ignored.

The fifth and final subgroup of aspects of MR instruction includes instances of *positive infidelity* (Cordray & Hulleman, 2009). These aspects reflect ideas drawn from

research on mathematics teaching and from studies in educational psychology that focus on how to support children's learning of mathematics with conceptual understanding. However, they are prohibited by the MR model because they require more direct involvement on the tutor's part than what is prescribed by the MR model. Thus, I view these aspects as potential local adaptations that contradict MR's program theory, but could potentially represent possible improvements to the model. They include: (a) re-voicing a student's (often incomplete) explanation to highlight particular mathematical ideas or to introduce mathematics vocabulary (Franke, Kazemi, & Battey, 2007; O'Connor & Michaels, 1993); (b) asking the student to solve a task (s)he has just solved in a different way, so that the student has an opportunity to make different mathematical connections or to represent mathematical relationships in a different way (Carpenter & Lehrer, 1999: NCTM, 2000); and (c) asking the student to compare alternative strategies and explain why they work so that they students has the opportunity to make connections between various strategies and mathematical ideas (Carpenter & Lehrer, 1999; Rittle-Johnson & Star; 2007). Some aspects in the nature of instruction subgroup described above are explicitly prohibited by the MR model, whereas the positive infidelity components are only implicitly prohibited. Nonetheless, they comprise an additional subgroup that pertains to quality of delivery, because they reflect process aspects of tutors' delivery of MR.

The positive infidelity aspects address an important issue that has come to the fore in ongoing debates concerning fidelity and adaptation (Blakely, Mayer, Gottschalk, Schmitt, et al., 1987). As summarized by Dane and Schneider (1998), a number of researchers have argued that any adaptations to a program model undermine efforts to

determine the program's efficacy. However, other researchers have argued that modifications to accommodate local needs can improve the success of a program. In including the positive infidelity components in the fidelity coding scheme for the MR program, I have elected to treat the issue empirically (e.g., Penuel & Means, 2004). That is, I have drawn on research on mathematics teaching to supplement MR's program theory with instructional practices that have been shown to effectively support students' mathematics learning (hence the term *positive* infidelity) in order to test whether such adaptations add or detract to the effects of MR tutoring.

As I have indicated in this section, the five subgroups of fidelity aspects account for four of the fidelity criteria identified in the literature: exposure/duration, adherence, quality of delivery, and participant responsiveness. One criterion is therefore unaccounted for: program differentiation. Dane and Schneider (1998) defined program differentiation as "a manipulation check that is performed to safeguard against the diffusion of treatments, that is, to ensure that the subjects in each experimental condition received only planned interventions" (p. 43). The absence of detailed data on control students' mathematics learning opportunities is likely a limitation of the study. Ideally, we would have assessed the extent to which regular classroom instruction resembled that of MR tutoring in order to calculate the "achieved relative strength" of the intervention (Cordray & Hulleman, 2009). That is, we would have compared the achieved strength of treatment delivered in tutoring (as compared to the MR model) to the "treatment" received by those students not in tutoring.

Nevertheless, program differentiation as defined by Dane and Schneider is not as crucial with respect to FOI of the MR program as it is with many other evaluations for at

least four reasons. First, MR is a pull-out program that *supplements* regular classroom instruction. The evaluation study therefore compared outcomes of students who received the supplement with those of students who did not receive the supplement (rather than tutoring versus 'business as usual'). Second, tutoring was delivered by tutors who were not the students' regular classroom teachers, and the tutors were told explicitly not to share what they had learned in their training with classroom teachers, so the likelihood of diffusion was nearly zero. Third, all sites were able to support the implementation of the program structurally, in that every school successfully designated a physical space dedicated to one-to-one instruction, and necessary recording equipment was provided to and used by all tutors. Fourth, none of the students who remained on the wait list received MR tutoring (or any other mathematics intervention), and none of the students who received MR tutoring received any other mathematics intervention.

Developing Instruments for Assessing Fidelity of Implementation

Continuing with the guidelines provided by O'Donnell (2008) and Nelson and colleagues (2010) outlined above, the next step was to develop a coding scheme for capturing all of the components shown in Figure 5, and to create coding instruments. Operationalizing key aspects of the program theory and developing appropriate indices required meeting the challenge of bringing developers' and researchers' perspectives to a consensus and making explicit what was previously largely implicit. As part of this process, I presented early versions of the coding scheme at two annual USMRC conferences in order to solicit feedback from both program developers and practitioners. Their reactions were, in general, supportive but critical. The development of the coding

scheme culminated in a three-day consultation with program developers in late 2008, during which a consensus regarding MR's core components and constructs that should be assessed and a means for codifying these in fidelity assessment instruments was finally reached. It is not surprising that this process was long and somewhat arduous. Although the program developers had a relatively well-articulated theory, particular aspects of the program remained largely implicit and no measures had previously been created. Over the first few weeks of 2009, another member of the research team and I finalized the instruments through an iterative refinement process, based on multiple rounds of independent video coding, discussion and further operationalization, eventually establishing adequate (90%) agreement. We then created computerized versions of the instruments for use by coders, using the database software, Filemaker®.

The instrument for coding the initial assessment is relatively straightforward; it was designed to be used while observing (on videotape) the tutor conducting the initial assessment. As described above and listed in Figure 5, in addition to confirming that the initial assessments were indeed administered, coders apply three types of codes to assess the extent to which that administration aligns with developers' intentions. These include: (a) noting tutors' administration errors, both major and minor (as defined above), on each of 17 sections of the assessments; (b) assessing whether, for each of the five (of six) aspects of the MR Learning Framework covered by the initial assessment, the initial assessment provides sufficient information about the child's thinking to assign a stage or level on the Framework; and (c) whether the tutor's profile assignment was correct.

The instruments for coding MR instruction are more complicated. Codes are applied at three levels: 1) lesson (a single tutoring session, typically lasting 25 minutes);

2) teaching procedure (a set of related tasks using the same instructional materials, or "setting"); and 3) task (any question posed by the tutor that requires the student to produce a numerical answer or solve a number-based problem). Figure 6 shows the coding scheme for aspects of MR instruction and indicates the level at which codes are applied.

At the lesson level, in addition to noting the length of the lesson (in total minutes), the codes capture whether the tutor's assignment of the student's profile on the MR Learning Framework upon entering a particular lesson was correct. Then, at the teaching procedure level, in addition to noting the total time spent on each procedure, the codes capture whether the choice of procedure aligned with the student's current profile (as indicated by the MR Instructional Framework). To answer these questions, coders must know the student's profile at the outset of any given tutoring session. Rather than relying on the tutor's assessment (as recorded on lesson plans), which may or not be correct, coders must view (up to three) previous tutoring sessions to identify evidence of the level of sophistication of the student's thinking on the MR Learning Framework. The process for doing so is similar to using video recordings of the initial assessments to assign profiles, but differs in that the tutor is not following a protocol designed to produce the necessary information to assign a profile. The coder's task is therefore similar to that of the tutors; the coder must use students' responses to tasks posed by the tutors in previous sessions to assess students' current number knowledge and arithmetical strategies and assign a profile. This, of course, requires that coders receive training on the MR frameworks and on interpreting students' responses during tutoring sessions I describe this training below.

		Elements	Definition	Level of Coding	Scoring	Operationalization
Supporting child's cognition	**ZPD**	Task challenge	Extent to which task afforded student opportunity to struggle meaningfully with mathematics; it was a "genuine problem." This rates what the task *became* if there was adjustment.	Task	Too Easy (E)	Task requires no struggle; student has automatized procedure and/or answers with certitude
					Genuine Problem (GP)	Neither too easy nor too difficult. Could have *started* as too difficult and become a genuine problem with adjustment of task challenge (see below).
					Too Difficult (D)	Student clearly has no basis for which to understand/solve problem or reaches an impasse. Adjustment with behavior eliciting would warrant a code of too difficult.
		Teaching procedure-portrait alignment	Whether teaching procedure targets the progression from one Learning Framework stage/level to the next	Teaching procedure	Yes	Purpose of procedure matches those suggested in MR handbook/directionality chart in terms of targeted skills, number range, setting, etc.
					No	Purpose of teaching procedure does not match those in MR handbook/directionality chart.
	Ongoing Assessment	Adjustment of task challenge	After tutor poses task & student cannot answer correctly (particularly cases where the student is not using any strategy), whether tutor reduces level of difficulty of task (*within* a task)	Task	Yes	At impasse, tutor adjusted question, but still targets original answer.
					No	No impasse was reached; OR student reached impasse without a tutor response.
		Responsiveness to student thinking/ answer	Whether tutor's subsequent instructional choices are in response to student's strategies/thinking on previous strategy-focused task. (If two tasks are posed together, rate only the first task.)	Task	DNA	First task in a procedure that targets a different strand (C/G/W) of the Learning Framework than the previous procedure
					Yes	Previous task → New task E → more challenging GP → at least as challenging (or no more challenging if previous task was GP w/adjustment) D → less challenging
					No	E → no more challenging GP → less challenging (unless previous task was GP w/adjustment) D → at least as challenging
		Evidence of student strategy	Whether, based on observation of student solving task and/or response to tutor's explicit solicitation, student thinking/strategy was evident	Task	Yes	With moderate level of inference it is clear *how* student came to a solution. Includes instances of automatization (unless it is revealed student has no conceptual foundation underlying the answer).
					No	Process by which student came to solution is not at all clear
		Correct Profile/ Portrait	Whether external assessment agrees with tutor's profile/portrait of student	Lesson	Yes	Stage/level assignments are equal
					No	Stage/level assignments not equal

Nature of Instruction	Post-task wait-time	Whether tutor allows sufficient time for student to think/problem solve/answer	Task	Yes	Student reaches solution or impasse without tutor interruption
				No	Tutor interrupts student thinking/solving
	Child Checking	When tutor asks/allows student to check his/her (last) answer (typically with a reduction in difficulty)	Task	DNA	Tutor did not ask/allow student to check answer or tutor does the work of checking
				Incorrect	Asks/allows student to check answer after incorrect student response
				Correct	Asks/allows student to check answer after correct student response
	Solicitation of student strategy	When tutor explicitly asks student to explain strategy/thinking	Task	DNA	Tutor did not solicit a strategy or solicited a confirmation (dichotomous question)
				Incorrect	Solicits after incorrect student response
				Correct	Solicits after correct student response
	Method demonstration	Whether tutor (un)intentionally demonstrates the method for how to solve a task	Task	Yes	Tutor indicates a strategy/method for solving task *before* or *after* task. This includes instances when the tutor solves the task for/with the student when the student's thinking was not evident (sometimes when tutor solicits student confirmation of strategy).
				No	Tutor does not indicate a strategy/ method for solving task *before* or *after* task
	Behavior eliciting	Whether tutor directs the student to solve a task (or at least a step) in a particular way; or emphasizes a particular solution method or response	Task	Yes	Tutor indicates a strategy/method for solving task *during* task (i.e., after student has begun thinking/solving)
				No	Tutor does not indicate a strategy/ method for solving task *during* task
Positive Infidelity	Re-voicing	Whether tutor re-voices student's strategy	Task	Yes	After task, tutor highlights, clarifies or represents student strategy. Instances when the tutor enlists the student's help are likely new tasks.
				No	Tutor does not highlight, clarify or represent student strategy. Note that re-voicing differs from Method Demo in that the student must have first produced his/her own method.
	Different strategy	Whether tutor asks student to solve problem in a different way	Task	Yes	After student successfully solves task, tutor asks student to solve the same task in a different way.
				No	Tutor does not ask student to solve in a different way (does not include typical MR instruction such as asking student for a different finger pattern)
	Compare strategies	Whether tutor's questions encourage student to examine the mathematical similarities and differences among two or more strategies	Task	Yes	After student successfully solves task, tutor asks student to compare his/her strategy to another strategy (likely one the student has used before).This does not include instances when the tutor demonstrates a new method without asking student to compare it to his/her own.
				No	Tutor does not ask student to compare his/her strategy to another strategy.

Figure 6. Coding scheme for assessing FOI of MR instruction.

Once the coder has determined the student's profile coming into a tutoring session, the coder then determines 1) whether the tutor's profile of the student's thinking matches their own independent assessment, and 2) whether the tutor's choice of procedures matches the child's placement on the MR Learning Framework. That is, did the tutor's choice of procedures align with what the MR Instructional Framework suggested? Tutors frequently utilized procedures that are described in the MR handbook. However, when they incorporated procedures from other sources (as is encouraged by the MR model), coders have to locate those procedures within the Instructional Framework. In making this decision, coders should consider three types of information: 1) the aspect of the framework on which the procedure focuses, 2) the number range targeted by the tasks within the procedure, and 3) the types of materials provided to the student.

Regarding the first consideration, coders have to determine which of the six aspects of the MR Learning Framework was being targeted, which in turn requires identifying the instructional goal of the procedure. For those who are very familiar with the MR frameworks, it is generally not difficult to determine whether a set of tasks is aimed at supporting a student in (a) developing more sophisticated arithmetical strategies, (b) voicing the number word sequences in correct order, (c) correctly naming numerals, or (d) counting by groups of ten. For the second and third considerations—number range and materials—coders use the descriptions of these characteristics of procedures in the MR handbook to find the closest match for any type of task the tutor used that was not, itself, specified in the handbook. Coders then apply the same criteria for determining whether the procedure was aligned with the kinds of tasks suggested by the MR Learning Framework.

Hiring and Training Coders

Five coders, each with experience in either elementary classroom instruction or video coding (or both), were hired and received two kinds of training: a five-day session led by two MR experts[2] on how to do Math Recovery (similar to the training tutors in the study received), and four days of training on using the coding instruments described above (led by members of the evaluation team). The MR training included (a) an introduction to the guiding principles of the program; (b) an examination of the distinctions between levels on the MR Learning Framework; (c) a trip to a local school to administer the MR assessment to first-grade students; (d) an introduction to the materials typically used in MR instruction; and (e) direction on coordinating the Learning Framework with the Instructional Framework.

It was essential to provide coders with extensive training on the program itself because much of what is required of coders closely resembles what MR tutors must do. For example, just like tutors, coders must be able to (a) use students' responses to mathematical tasks to assign a profile on the MR Learning Framework; (b) determine which instructional procedures would be appropriate given a student's current profile; and (c) determine whether each task represents a "genuine problem" for the student. Additionally, coders must be able to judge whether a tutor's actions when introducing tasks and responding to a student's solution align with the MR model. Our assumption was that their assessments of fidelity would be more likely to faithfully adhere to the

[2] These were the same two MR experts who provided the training to the participating tutors.

intent of the coding scheme if they understood MR's underlying theory and used its fundamental tools (i.e., frameworks and instructional materials).

The initial four-day coding training, led by myself and another member of the research team, included (a) an introduction to the operationalizations of the core implementation components of MR, including the coding instruments; (b) multiple rounds of collective video coding, in the course of which we discussed coding decisions; and (c) initial independent coding with group discussion immediately following. The last phase of training included (d) completely independent coding for which percent agreement was determined until an adequate level of agreement was reached consistently.

Throughout the final phase, which lasted four weeks, my colleague from the research team and I met weekly with the coders to further refine, define and operationalize the aspects of MR that they were attempting to code. Thus, early on, coders' feedback was important in improving the feasibility of MR fidelity assessment. However, after four weeks of refinement work, agreement percentages plateaued at an inadequate level, largely due to differences in how coders 'chunked' the lessons they were coding (e.g., was it one big task, or two small tasks?) Therefore, my colleague from the research team and I identified a representative aspect of the MR Instructional Framework about which coders' structural decisions had consistently agreed and for which all codes would remain relevant. Of the six aspects included in the MR Learning Framework, two of them (Stages of Early Arithmetical Learning, and Tens and Ones) represent the heart of the theory underlying the MR program. Although lessons typically include practice on other aspects such as number word sequences or numeral identification, it is these two aspects that pertain directly to the unique aspects of MR

listed above (and on which the MR Proximal outcome measure—designed in collaboration with MR developers—focuses primarily). The decision was therefore made to restrict the coding to tasks that aimed at supporting students in developing more sophisticated arithmetical *strategies*. This decision rendered the fidelity assessment process more tractable without sacrificing any of the core implementation components. The resulting approach is consistent with Mowbray and colleagues' (2003) notions concerning process-oriented components of interventions, which are more difficult to reliably assess than structural features.

> In many instances, the elements of a fidelity measure serve, in effect, as indicators of the model's design and operations—key program features that relate strongly to positive outcomes for those served—but do not necessarily include all such features, nor any features in the depth suggested by a fully explicated program theory. Indicators are selected, then, on an empirical basis (relationship with outcomes), and also because they are reliable and easy to measure (Mowbray et al., 2003; p. 330).

After limiting our within-lesson fidelity assessment to the two aspects of the Learning Framework named above, coders were able to achieve an observer-agreement percentage of at least 80% on all classes of codes combined during the final training/refinement phase. (Disaggregated, code-specific agreement rates are reported in Table 2 and discussed in the section on reliability.)

Sampling

The fourth guideline identified by O'Donnell (2008) for conducting fidelity studies concerns sampling within the study in order to generalize fidelity findings to the study population. In this section I describe my frames for sampling at two levels:

sampling students within tutors and sampling lessons (i.e., tutoring sessions) within students.

As I have indicated, all assessment and tutoring sessions conducted by the 18 tutors were video-recorded in line with standard MR practice. This resulted in more than 5000 hours of video and a practical necessity of coding only a sample of the data. In order to ensure that every tutor's practices across both years of the study are represented in FOI analyses, I randomly selected one student (of three) from each cycle per tutor. This resulted in a total of 107 students across all 18 tutors and all six cycles during the two-year study (one tutor conducted only five of the six cycles of instruction due to a maternity leave).

For each student selected, one coder (or two coders, if, as explained in the next subsection, the student's videos were selected for ongoing reliability checks) assessed the fidelity with which the initial assessment and 12 instructional lessons were conducted. This large sample of lessons was coded in order to produce an adequate number of data points for various secondary analyses, including a generalizability study to determine the minimum number of observations necessary to produce a stable estimate of a tutor's fidelity to MR within a cycle. To select the lessons for coding, I divided the total number of tutoring sessions for each student into six equal blocks and randomly selected two lessons from each block. This resulted in a total 107 assessments and 1,284 tutoring sessions conducted with 107 students.

Video-recorded lessons were assigned to coders randomly rather than chronologically. For example, a given coder might begin with a student from the sixth and final tutoring cycle, and follow that with a student from the very first tutoring cycle.

In assigning coding assignments in this manner, I avoided confounding a change in tutors' fidelity to the MR model over time with coder drift (although, as explained below, procedures were implemented to minimize coder drift).

Reliability and Validity

I and another member of the research team assessed coder agreement (sometimes called "inter-rater reliability") throughout the coding process. Of the 107 student cycles of tutoring selected for fidelity coding, 21 (approximately 20%) were randomly selected to be double-coded in order to assess coder agreement. These 21 cycles comprised 21 initial assessments and approximately 252 instructional sessions. Coders remained blind to this selection until after they had completed their individual coding. I scheduled double coding regularly throughout the coding process so that each coder was involved in at least one double-coded case every two weeks.

Table 2 includes disaggregated coder agreement rates for all fidelity indicators, including both those for the initial assessment and those for instructional sessions. For the initial assessment, we assessed coder agreement by conducting a simple cell-by-cell comparison of their codes. Table 2 lists the percentages of exact agreement for each indicator. Rates of exact agreement were above 70% for 20 of the 27 codes that pertained to the initial assessment.

For instructional sessions, coder agreement was assessed at several levels, because of the nested structure of the coding scheme. Because the coder first must decide whether or not there are strategy-focused activities to be coded within a given lesson, we first checked coder agreement at that level. Coders agreed on whether an instructional

session should have been coded (i.e., included at least one strategy-focused activity) 90.5% of the time. (On average, 7 of 12 selected instructional sessions included strategy-focused activities, so any given lesson was only slightly more likely to warrant coding than not.)

Table 2 lists four calculations of coder agreement for other indicators, which were coded at the task-level. The first column lists correlations between pairs of coders for aggregated versions of each indicator (calculated as described below in the section on variable construction). The second reports percentages of instances of double-coding in which the difference between coders' aggregated scores for a particular indicator were less than one standard deviation (calculated for each indicator on the entire set of fidelity data). When disagreements occurred at the lesson level, we asked coders to come to consensus about whether there was a strategy-focused activity within a particular lesson, and make the appropriate coding adjustments (without talking about the coding details at the other levels). Once the two coders' sets of codes matched structurally at the lesson level, we calculated an overall agreement percentage by comparing codes cell-by-cell. Once we had determined the coder agreement percentage, coders resolved all disagreements (at the task level) and arrived at a consensus.

The remaining columns of Table 2 report agreement rates after such structural disagreement had been resolved. The third column lists straightforward, cell-by-cell agreement rates. As listed in the fourth column, another member of the evaluation team and I also calculated rates of agreement when limited to the tasks that had clearly been coded by both coders based on coders' brief descriptions of the tasks and students' responses to those tasks. Often, it was difficult to align coders' results in order to identify

Table 2

Coder agreement rates by fidelity indicator

Initial assessment (n=21)			**Instructional sessions (n=21)**				
Initial Assessment Indicators		Coder agreement	Instructional Session Indicators	Aggregated scores before structural disagreements resolved		Agreement After structural disagreements resolved	
				1) Correlations between coder pairs' scores	2) % within 1 SD	3) % agreement	4) % when limited to matched tasks
Type of Error Assessment: *Major*, *Minor*, or *None*	**1. Forward Number Sequence (FNWS)**	**0.95**	**Inclusion of strategy-focused activity**	(not calculated)	(not calculated)	0.78	n/a
	2. Number Word After	**0.73**	**Time spent on strategy-focused activities**	(not calculated)	(not calculated)	0.76#	n/a
	3. Numeral Identification	**0.82**	**Task challenge**	0.22	.76	0.63	0.79
	4. Numeral Recognition	**1.0**	Teaching procedure-portrait alignment	(not calculated)	(not calculated)	0.50	0.62
	5. Backward Number Word Sequence (BNWS)	**0.77**	**Adjustment of task challenge**	0.65***	1.0	0.75	0.93
	6. Number Word Before	0.68	Responsiveness to student thinking	-0.06	.62	0.57	0.71
	7. Sequencing Numerals	0.59	**Evidence of student strategy**	0.72***	.71	0.59	0.74
	8. Additive Tasks	0.55	**Post-task wait-time**	.70**	.76	0.71	0.88
	9. Subtractive Tasks	0.41	**Child Checking**	0.73***	.90	0.72	0.89
	10. Subitizing and Spatial Patterns	**0.77**	**Solicitation of student strategy**	0.88***	.95	0.75	0.94
	11. Finger Patterns 1 to 5	**0.82**	**Method demonstration**	0.42 (p=.05)	.76	0.71	0.88
	12. Finger Patterns 6 t o10	**0.91**	**Behavior eliciting**	0.84***	.95	0.72	0.90
	13. Five Frame Patterns	**0.82**	**Re-voicing**	0.86***	.95	0.77	0.96
	14. Five-wise Patterns	**0.95**	**Different strategy**	0.96***	.95	0.79	0.99
	15. Pair-wise Patterns	0.59	**Compare strategies**	0.51*	.81	0.79	0.99
	16. Combining to Make Five	**0.73**	***p<.001 ** p < .01 * p < .05 **#**Percent of instances in which both coders agreed on total number of minutes spent on strategy-focused activities within *p* minutes, where *p* is the number of teaching procedures coded for that student (which ranges from 3 to 26, with a mean of 14.5 and standard deviation of 6.30).				
	17. Combining to Make Ten	**0.82**					
Sufficient information revealed	**Stage of Early Arithmetical Learning (SEAL)**	**0.91**					
	FNWS	**0.86**					
	BNWS	**0.82**					
	Numeral identification	**0.91**					
	Structuring number	**0.95**					
Profile assigned by tutor correct	**SEAL**	**0.82**					
	FNWS	0.36					
	BNWS	0.64					
	Numeral ID	**0.77**					
	Structuring number	**0.86**					

exactly which tasks' codes should be compared between coders. Consequently, estimated rates of agreement were likely conservative; re-calculating them in this manner helps determine the extent to which the coders applied the coding scheme consistently to particular tasks.

In total, of 42 fidelity codes, our coders applied 33 with at least 70% agreement (which, for task-level instructional session codes, I defined as having at least 0.7 in at least two of the columns). However, for task challenge and evidence of student strategy, the reliability estimates were modest. As discussed below, I did not include in my analyses any of the nine indicators for which coder agreement was less than 70%.

Because coders discussed their coding at multiple stages to reach a consensus, the data they generated are not entirely independent. I therefore included an examination of potential rater effects in my analysis (the outcome of which suggested no significant differences between coders).

I conducted two analyses for purposes of validation. First, the coding was used to code the fidelity of tutoring in approximately 15 MR training video-recordings. Because the recordings are used in MR training as exemplars of high quality MR tutoring, the tutoring practices therein presumably aligned with the MR model. As will be shown with the full distributions of FOI scores below, the fidelity scores of the tutors in the training videos were generally high compared with study tutors' scores. Because the application of our instruments categorized the exemplary tutoring episodes as "high fidelity" enactments of MR, this is one confirmation of the validity of the coding scheme.

Second, following Mills and Ragan's (2000) recommendation of consulting with program developers, I submitted a subset of assessment and tutoring sessions to 12 MR

experts, who rated the tutoring practices based on their notions of high-quality MR practice. This subset consisted of eight initial assessment interviews, eight full lessons, and eight excerpts of lessons that included only strategy-focused activities selected to represent the full range of scores on indices of implementation fidelity as determined by my coding schemes. I randomly assigned six of each type of video to each of the 12 MR experts and asked them to 1) rank, from highest to lowest, the extent to which the tutors' enacted MR as intended, and 2) indicate in which of four categories they would place each video: *excellent*, *good*, *fair* or *poor* (the score sheet supplied to the MR experts is included as Appendix A). The decision to assign only six rather than all eight of each type of video to each expert was to make the rating task more manageable, and to prevent them from assuming that they should assign two videos per each of the four categories. Each video recording was labeled with a pseudonym for reference, and the MR experts remained blind to both the research team's instruments and assessment criteria, and each other's rating decisions until after they had completed their ratings.

After collecting all 12 MR experts' ratings, I calculated the Spearman's rank correlation coefficient both between expert raters and between the raters' average rankings and the scores determined by the fidelity coding. Table 3 shows the results of these calculations. For pairs of MR experts, mean correlations for assessments, full lessons and excerpted lessons were 0.59, 0.24, and 0.69, respectively, suggesting modest reliability among experts' ranking for assessment and excerpted lessons and low reliability for full lessons. The rank correlations between experts' average rankings of assessments, full lessons and excerpted lessons (which I determined by calculating an average ranking for each video recording within each type of video) and the rankings

determined by fidelity scores, were -0.07, 0.43, and -.05, respectively, suggesting a low level of agreement between fidelity scores and experts rankings for assessments and excerpted lessons, and a modest level of agreement for full lessons. However, if the assessment video recording that was ranked last based on fidelity scores, which experts consistently ranked among the top two (which could have been an unfortunate anomaly attributable to measurement error among those videos selected for the validity assessment), was removed from the calculation, the rank correlation was 0.39, a level of agreement comparable with that of full lessons.

Table 3
Results of expert raters' video categorizing and ranking

Fidelity study rank	Initial Assessments		Full Lessons		Lesson Excerpts	
	Raters' avg category	Raters' avg rank	Raters' avg category	Raters' avg rank	Raters' avg category	Raters' avg rank
1	fair	2.67	fair	3.44	fair	4.13
2	fair	3.44	fair	1.89	fair	3.33
3	fair	3.22	fair	2.89	poor	5.22
4	poor	5.33	poor	5.11	good	1.11
5	good	2.44	fair	3.33	fair	3.00
6	poor	5.22	fair	3.22	fair	3.33
7	poor	4.44	poor	4.44	good	2.22
8	good	1.22	fair	3.67	poor	5.78
Spearman's rho		**-0.07**[a]		**0.43**		**-0.05**

Note. Possible categories included *excellent*, *good*, *fair*, and *poor*. Spearman's rho values represent rank correlations between the fidelity study rankings and raters' average rankings for each type of video.
[a]Spearman's rho for assessments increases to 0.39 if the video ranked as number 8 by fidelity scores is removed from the calculation.

This test of instrument validity was very stringent and limited by a number of factors. First, Spearman rank correlations are not a highly stable estimate of agreement in this case because of the small number of rankings (a comparison of between 6 and 8 video recordings for each pair of expert raters, and between 12 between fidelity scores and expert rankings). Second, there was likely a significant contrast between the types of

schemes applied by the two groups (trained fidelity coders and MR experts). On the one hand, the fidelity scores were produced by applying codes systematically and using a weighted linear scheme. On the other, experts' ratings were likely produced from configuration schemes, with decisions based on a series of 'red flags' (i.e., particular tutor behaviors that the expert considers to be a violation of MR program expectations). Therefore, it is not surprising that agreement between the two would be modest. Considering this limitation, the correlation between fidelity scores and average MR experts' rankings for full lessons (0.43), which was higher than the average correlations between pairs of expert raters, is somewhat encouraging, since fidelity scores were determined based on the portions of lessons that focused on strategy-based activities (like the excerpted lessons ranked by MR experts).

Table 4
Frequencies of expert raters' video categorizations

Fidelity study rank	Initial Assessments				Full Lessons				Lesson Excerpts			
	exc.	good	fair	poor	exc.	good	fair	poor	exc.	good	fair	poor
1	0	2	7	0	1	1	4	3	0	2	2	4
2	2	0	3	4	2	1	3	2	1	2	6	0
3	0	3	5	1	0	2	3	4	0	0	3	6
4	0	0	2	7	0	0	3	6	4	2	3	0
5	2	3	3	1	1	2	1	5	1	4	3	2
6	0	0	1	8	0	3	2	4	2	2	2	3
7	0	1	2	6	0	1	3	6	1	5	2	1
8	5	2	2	0	1	1	2	5	0	0	0	9
Percent	0.13	0.15	0.35	0.38	0.07	0.15	0.29	0.49	0.13	0.24	0.29	0.35

Last, the validity test was limited by a significant lack of agreement among expert raters (which is perhaps an important finding in itself). Table 4 shows the results of expert raters' categorizations of the video recordings. As can be seen, considerable

variation existed across expert raters with respect to a number of video recordings. Often, while some expert raters categorized a video as "excellent" or "good," others categorized the same video as "fair" or "poor." This variation suggests that expert raters' configuration schemes were fairly idiosyncratic. An informal analysis of the notes submitted by half of the expert raters indicates that even in some of the cases in which raters agreed on rankings or categorizations, they stated different criteria for their decisions.

Analysis

In this section, I divide the final step of O'Donnell's (2008) and Nelson et al.'s (2010) six guidelines discussed in Chapter 2 into three subsections. I first detail the creation and properties of the FOI indices for use in my analyses, including variable construction, distribution and variance composition. Then, I address the possibility of rater effect. Finally, I step through my research questions, including descriptions of the models I employed to answer those questions.

FOI variables

Over an entire set of MR teaching procedures coded for each randomly selected student, I aggregated the data generated by the coders to create variables for using in my FOI analyses. In this subsection, I first describe my original conceptions of those variables, and then report ways in which they were adjusted to account for the inadequate

levels of reliability among some of the fidelity indicators and the correlational structures among indicators intended to be aggregated into single scales.

Figure 7 includes my original conception of potential FOI variables. Those in bold type represent the predictors I intended for my models; many were comprised of the variables listed in plain type. In total, I intended to create four variables representing fidelity of tutors' uses of the initial MR assessments (*minerr*, *majerr*, *infototal*, and *profperc*); three indices of exposure/duration (*lssn_no*, *meanlesstime*, *avgsealtime*); and four variables representing the fidelity of tutors' MR tutoring practices (*assmt*, *zpd*, *noi*, and *posinf*). Of the variables pertaining to tutoring, those accounting for ongoing assessment (*assmt*) and instruction delivered within the student's ZPD (*zpd*) were to represent the two unique and essential components of the MR program; the variable pertaining to nature of instruction (*noi*) was intended to include the essential but not unique components; and the variable pertaining to positive infidelity (*posinf*) was to include potentially effective adaptations that are inconsistent with the MR model (as described above).

Because the reliability levels for some indicators reported above were inadequate, I had to modify 5 of the 11 variables that I had originally planned to use. The first three pertain to the initial assessment. Because 5 of the 17 subsections of the assessments were not coded reliably, I limited both error variables (*minerr* and *majerr*) to the 12 that were coded with sufficient reliability. Additionally, because students' levels on two aspects of the MR Learning Framework, FNWS and BNWS, were not coded reliably, the variable that captures tutors' rates of accuracy in assigning initial profiles for students (*profperc*)

	FOI Variable	Definition	Values	Aggregate:
INITIAL ASSESSMENT	**minerr**	**Total number of minor errors committed by tutor in administering the initial assessment.**	**[0, 17] (discrete)**	
	majerr	**Total number of major errors committed by tutor in administering the initial assessment.**	**[0, 17] (discrete)**	
	infototal	**Total (out of 5) of aspects of the MR Learning Framework for which tutor used the initial assessment to generate sufficient information to assign a profile stage/level**	**[0, 5] (discrete)**	
	profperc	**Percentage of 5 aspects of MR Learning Framework on which the tutor's assignment of student profile matches that of the coder**	**{0, 0.2, 0.4, 0.6, 0.8, 1}**	**Average of 5 dummy codes for each of the aspects of the Learning Framework assessed by the initial MR assessments**
LESSONS	**lssn_no**	**Number of lessons provided to student**	**[1,…] (discrete)**	
	meanlesstime	**Average length of lessons in minutes**	**[0,…] (continuous)**	
	sealtaoadd	Percentage of lessons coded in which SEAL and/or Tens & Ones (strategy-focused) procedures were employed	[0, 1] (continuous)	Average of dummy codes from each lesson (typically 12)
	avgsealtime	**Average number of minutes spent on strategy-focused teaching procedures per lesson**	**[0,…] (continuous)**	
	adj	Percentage of tasks that tutor reduced level of difficulty of task (*within* a task) when initial level of task challenge was too difficult	[0, 1] (continuous)	
	evid	Percentage of total number of tasks posed to student in which student thinking/strategy was evident	[0, 1] (continuous)	
	resp	Percentage of instances in which the tutor's task choice is responsive to student's strategies/thinking on previous strategy-focused task (i.e., student's thinking/strategy was evident in the last task, and current task is a genuine problem or closer to being a genuine problem than the last)	[0, 1] (continuous)	
	assmt	**Tutor's average rate of 'ongoing assessment'**	**[0, 1] (continuous)**	**Average of adj and resp**
	tcgp	Percentage of total number of tasks posed to student that were genuine problems	[0, 1] (continuous)	
	alignperc	Percentage of procedures used by tutor that aligned with (coder's) profile of student according to the MR Instructional Framework	[0, 1] (continuous)	
	zpd	**Tutor's average rate of providing instruction in the student's ZPD (as characterized by MR)**	**[0, 1] (continuous)**	**Average of tcgp and alignperc**
	wait	Percentage of total number of tasks posed to student after which tutor allowed sufficient time for student to think/problem solve/answer	[0, 1] (continuous)	
	behav	Percentage of total number of tasks posed to student in which tutor directed the student to solve a task (or at least a step) in a particular way; or emphasizes a particular solution method or response	[0, 1] (continuous)	
	demo	Percentage of total number of tasks posed to student before/after which tutor (un)intentionally demonstrates the method for how to solve a task	[0, 1] (continuous)	
	check	When, on average, tutor asks/allows student to check his/her (last) answer (typically with a reduction in difficulty)	[-1, 1] (continuous)	
	solic	When, on average, tutor typically explicitly asks student to explain strategy/thinking	[-1, 1] (continuous)	
	noi	**Tutor's average rate of employing MR-appropriate instructional moves**	**[0, 1] (continuous)**	**Average of wait, (1 – behav), (1 – demo), (1- |check|) and (1 - |solic|)**
	revoice	Percentage of total number of tasks posed to student after which tutor re-voiced student's strategy	[0, 1] (continuous)	
	diff	Percentage of total number of tasks posed to student after which tutor asked student to solve problem in a different way	[0, 1] (continuous)	
	compare	Percentage of total number of tasks posed to student after which tutor's questions encouraged student to examine the mathematical similarities and differences among two or more strategies	[0, 1] (continuous)	
	posinf	**Tutor's average rate of employing positive infidelity moves**	**[0, 1] (continuous)**	**Average of revoice, diff and compare**

Figure 7. Original conception of variables to be created for FOI analyses.

was limited to three aspects of the MR Learning Framework: SEAL, Numeral Identification, and Structuring Number. I weighted SEAL twice the other two aspects of the MR Learning Framework because SEAL represents the core of the models of children's learning in early number on which MR draws, around which the other aspects of the Learning and Instructional Frameworks are built (indeed, the word "stages" is reserved for only SEAL; all other aspects of the MR Learning Framework consist of "levels").

The remaining two reliability-related modifications pertain to the instructional sessions. First, because the percentage of instances in which the tutors' task choices were responsive to students' thinking on previous strategy-focused tasks (*resp*) was not coded reliably, the variable I had intended to use to capture tutors' fidelity to "ongoing assessment" (*assmt*) had to be limited to the percentage of tasks for which tutors reduced the level of difficulty when the initial level of task challenge was too difficult (*adj*). Second, similar to the *profperc* variable for the initial assessment, the alignment of tutors' teaching procedures and their students' MR Learning Framework profiles (according to the MR Instructional Framework) was not coded with sufficient reliability. Therefore, it had to be dropped from the variable that was intended to capture tutors' success at delivering instruction within students' zones of proximal development (*zpd*), instead limiting that variable to the percentage of tasks posed by tutors that were "genuine problems" for students (*tcgp*).

After limiting the fidelity indicators to those coded with adequate reliability, 17 remained. As indicated in Figure 7, most of the variables are unweighted averages across all of the tasks posed to the student. For example, *meanlesstime* is the mean length in

minutes of tutoring sessions a student received, calculated by averaging the lengths of six sessions chosen from the beginning, middle and end of a tutoring cycle. Likewise, the variable *tcgp* represents the ratio of tasks that were coded as "genuine problems" to the total number of tasks coded for a student from across his/her entire tutoring cycle. The *adj* variable is also a percentage, similar to those mentioned above, but calculated conditionally. It accounts for the percentage of instances when, after posing a task that is too challenging, the tutor adjusts the level of difficulty of the task. It is calculated by dividing the total number of times the tutor adjusted the task difficulty by the total number of tasks posed to the student that were too difficult (which excludes those tasks that were genuine problems or too easy). Other variables are simple counts. For example, *lssn_no* indicates the total number of lessons a student received; *majerr* and *minerr* represent the total number of major and minor errors the tutor committed during the initial assessment.

A majority of the indices are continuous, ranging from 0 to 1, with the exception of those that that are simple counts or average lengths of time, *profperc* (the percentage of aspects on the MR Learning Framework on which the tutor's assignment of student profile matches the coder's—of which there were only three coded reliably), and two indices that were intended to be combined in the nature of instruction variable: *check* and *solic*. As explained above, indices of child checking (requiring the student to check his/her own solutions by employing a different strategy than the one employed to reach the original solution) and solicitation of student strategy (soliciting an explanation from the student after solving a task) do not indicate mere frequencies of particular actions. More important than the frequency with which a tutor asks a student to check her/his

work or explicitly asks the student to explain the strategy (s)he has used to solve a task, is whether there is a pattern to the tutor's uses of these moves. Instances of either of these tutor behaviors are coded as -1 if they occurred after an incorrect response from the student and 1 if they occurred after a correct response. Thus, when averaged across all coded tasks, these indices range from -1 to 1, with zero representing a 'perfect balance' between types of response.

As indicated in Figure 7, I originally intended to combine certain indicators to create four scales pertaining to instructional sessions: *zpd*, *assmt*, *noi*, and *posinf*. As already explained, because particular indicators were not coded with sufficient reliability, it was no longer feasible to create the first two, *zpd* and *assmt*,. Therefore, only two scales remained to be constructed: tutors' nature of instruction (*noi*) and instances of positive infidelity (*posinf*). In the next five paragraphs, I document my work in constructing those scales.

I determined the internal consistency (as indexed by Cronbach's alpha) of the two scales I intended to create by examining the inter-correlations among the indicators I intended to combine. However, these correlations were influenced by the clustering of data pairs within tutors. That is, the direction and magnitude of the relationships between pairs of indicators were likely related to patterns of fidelity to the MR model that varied by tutor. As show in Appendix B, fidelity to particular indicators of nature of instruction increased for some tutors over the six cycles of tutoring, while it decreased for others. Consequently, failure to take account of tutor clusters when examining inter-correlations of fidelity indicators would result in under-estimated alphas for both the *noi* and *posinf* variables. Therefore, I divided tutors into two groups—those whose fidelity to the nature

of instruction indicators increased over time (10 tutors, with a total of 60 time points) and those whose fidelity decreased (8 tutors, with a total of 47 time points), and calculated alphas for both the *noi* and *posinf* variables separately for each group.

I intended to create the nature of instruction variable by combining five indicators: *wait* (the percentage of total number of tasks posed to student after which the tutor allowed sufficient time for the student to think/problem solve/answer); *behav* (the percentage of the total number of tasks posed to the student in which the tutor elicited a particular behavior); and *demo* (the percentage of total number of tasks posed to the student before/after which the tutor (un)intentionally demonstrated the method for how to solve a task); as well as the two indicators described above, *check* and *solic*. Because *behav* and *demo* are both reverse-coded (with a frequency of 0 being most desirable) I subtracted their values from one. Because the direction of the *check* and *solic* indices matters less than their distance from the ideal zero, I subtracted their absolute values from one.

For the positive infidelity variable (*posinf*), I intended to combine three indicators, all of which are rates of frequency across all tasks, ranging from 0 to 1: *revoice* (re-voicing a student's explanation to highlight particular mathematical ideas or to introduce mathematics vocabulary); *diff* (asking the student to solve a task (s)he has just solved in a different way); and *compare* (asking the student to compare alternative strategies and explain why they work).

The results of my examination of internal consistency, both before and after diving tutors into two groups, suggested that for two of the nature of instruction indicators, *check* and *solic*, the relationship with the other indicators was not sufficiently

strong to warrant including them in the *noi* scale. After dropping those two indicators, the alpha for the *noi* scale was 0.73 for both groups of tutors (those whose fidelity decreased over time and those whose fidelity increased). For the *posinf* scale, the alpha for tutors whose fidelity decreased was 0.72, and 0.64 for those whose fidelity increased. These results are summarized in Table 5.

Table 5
Summary of alphas calculated for nature of instruction and positive infidelity variables before and after separating tutors into two groups

Standardized alphas	**Before dividing tutors into two groups**	**After dividing tutors into two groups**		**After dropping *check* and *solic***	
		FOI decreased (n=60)	FOI increased (n=47)	FOI decreased (n=60)	FOI increased (n=47)
noi	.5383	.6082	.7170	.7251	.7358
posinf	.6380	.7168	.6400		

Within the nature of instruction scale, behavior eliciting (*behav*) and demonstrating a method (*demo*) are similar types of prohibited tutoring moves, operationalized primarily by *when* they occur. Coders coded for method demonstration *before* a task was posed or *after* a student had completed a task; the behavior eliciting code was applied when, *during* a task (i.e., while the student was still working/thinking), the tutor intervened to elicit a particular response from the student. Therefore, in calculating the nature of instruction variable (*noi*), I weighted these equally, but only half as much as wait-time (*wait*), which represents a different type of tutor move. Thus, the *noi* variable was calculated as follows: *noi* = (2**wait* + *behav* + *demo*)/4. All three indicators included in the positive infidelity variable were weighted equally because they have been argued distinctly in the literature to have impacts on student learning. This yielded the following calculation for that scale: *posinf* = (*revoice* + *diff* + *compare*)/3.

Finally, because child checking (*check*) and solicitation of strategy (*solic*) were dropped from the nature of instruction scale, I created stand-alone variables for both. I used the absolute value of each index as an 'imbalance factor,' and weighted that factor by the frequency with which the tutor employed the move (the ratio of tasks when the move was used to the total number of tasks). My rationale for this, as stated above, is that the issue is not simply whether the tutor uses such moves. What is important, according to the MR model, is whether there is a pattern to the tutor's use of the moves. Hence, the weighted variables I created account for a repeated imbalance in when the moves are used (e.g., soliciting a strategy after incorrect responses more often than after correct responses, or vice versa).

Having described the construction of my variables, I now turn to two characteristics of the fidelity data: distribution and composition of variance. First, the FOI variables must have sufficient variation in order to link them to student outcomes. Table 6 lists the final variables I used in my analysis, including definitions and descriptives statistics for each. All variables yielded a unimodal distribution of values. However, in many cases, the distributions are somewhat skewed to the left, and in other cases the standard deviations are small. Although this is to be expected (since every tutor's implicit goal is to provide MR tutoring with fidelity), the distributions of some variables presented potential problems in linking them to student outcomes due to restrictions of range. These variables include three of the indicators pertaining to the initial assessment (*majerr*, *minerr*, and *infototal*); the average length of tutoring sessions (*meanlesstime*); and the three 'nature of instruction' indicators (*wait*, *behav*, and *demo*). I return to this issue in Chapter 4, when I report on the results of my analyses.

Table 6
Definitions and descriptive statistics of fidelity variables

Fidelity variable	Definition	Mean[a]	SD	Min	Max
minerr (out of 12) [c]	Total number of minor errors committed by tutor in administering the initial assessment.	1.07	1.12	0	6
majerr (out of 12) [c]	Total number of major errors committed by tutor in administering the initial assessment.	0.83	1.09	0	5
infototal (out of 5) [c]	Total (out of 5) of aspects of the MR Learning Framework for which tutor used the initial assessment to generate sufficient information to assign a profile stage/level	4.22	1.04	1	5
profperc (% out of 3)	Percentage of 5 aspects of MR Learning Framework on which the tutor's assignment of student profile matches that of the coder	0.69	0.31	0	1
lssn_no[c]	Number of lessons provided to student	32.37	7.72	3	52
meanlesstime[d]	Average length of lessons in minutes	25.04	3.00	16.92	31.38
avgsealtime[d]	Average number of minutes spent on strategy-focused teaching procedures per lesson	6.54	3.30	0.7	15.21
adj[b]	Percentage of tasks that tutor reduced level of difficulty of task (*within* a task) when initial level of task challenge was too difficult	0.83	0.25	0	1
tcgp[b]	Percentage of total number of tasks posed to student that were genuine problems	0.62	0.17	0.18	0.99
noi[b]	Tutor's average rate of employing MR-appropriate instructional moves: (2**wait* + *behav* + *demo*)/4	0.87	0.09	0.58	0.99
wait[b]	Percentage of total number of tasks posed to student after which tutor allowed sufficient time for student to think/problem solve/answer	0.87	0.10	0.53	1.0
behav[b]	Percentage of total number of tasks posed to student in which tutor directed the student to solve a task (or at least a step) in a particular way; or emphasizes a particular solution method or response	0.86	0.11	0.49	1.0
demo[b]	Percentage of total number of tasks posed to student before/after which tutor (un)intentionally demonstrates the method for how to solve a task	0.86	0.11	0.39	1.0
check[e]	When, on average, tutor asks/allows student to check his/her (last) answer (typically with a reduction in difficulty)	0.13	0.15	0.40	1.0
solic[e]	When, on average, tutor typically explicitly asks student to explain strategy/thinking	0.07	0.11	0.44	1.0
posinf[b]	Tutor's average rate of employing positive infidelity moves: (*revoice* + *diff* + *compare*)/3	0.01	0.02	0	0.12
revoice[b]	Percentage of total number of tasks posed to student after which tutor re-voiced student's strategy	0.03	0.05	0	0.26
diff[b]	Percentage of total number of tasks posed to student after which tutor asked student to solve problem in a different way	0.01	0.02	0	0.13
compare[b]	Percentage of total number of tasks posed to student after which tutor's questions encouraged student to examine the mathematical similarities and differences among two or more strategies	0.005	0.01	0	0.09

[a]n = 107 observations
[b]Mean frequency, [0-1]
[c]Total count
[d]Average minutes/lesson
[e]'Imbalance' factor weighted by frequency, [0-1]

Second, I examined the decomposition of variance within FOI indices to compare two possible ways of conducing my analyses. In the first way, only the 107 students whose tutoring sessions were coded to assess FOI would be included in the models. In the second way, each tutor's FOI scores based on one randomly selected student from a cycle would be applied to all three students in that cycle, and therefore all treatment students would be included in the models. To determine which of these alternatives was most appropriate, I used Stata's *loneway* command to determined where the majority of variance lies when the analysis is limited to only the 107 fidelity students: between or within tutors.

Table 7
Results of the analysis of variance of FOI variables

FOI variable	**SS**		**F**	**ICC**
	Between tutors	**Within tutors**		
minerr	23.41	109.13	1.12	0.02
majerr	37.44	89.53	2.19**	0.17
infototal	18.58	96.03	1.01	0.002
profperc	1.03	8.98	0.60	0.00
lssn_no	1246.68	5078.37	1.29	0.05
meanlesstime	596.19	355.64	8.78***	0.57
avgsealtime	590.51	563.42	5.49***	0.43
adj	1.40	5.32	1.37	0.06
tcgp	0.46	2.48	0.98	0.00
noi	0.25	0.53	2.47**	0.20
check	0.71	1.50	2.48**	0.20
solic	0.29	072	2.11*	0.16
posinf	0.19	0.04	2.60**	0.21

***p<.001
** p<.01
* p<.05

The results of my analysis of variance are displayed in Table 7. For a majority of the fidelity variables, the majority of the variance was within tutors, suggesting that tutors' fidelity of implementation varied too greatly from student to student to apply a

FOI score produced by coding tutoring sessions for one student to the other students in a tutor's cohort. Therefore, in my analyses of the relationship between FOI and student outcomes, I used data from only the 107 students for whom I calculated fidelity scores.

Rater effect

Having described my FOI variables and their properties, I now consider rater effects. As stated above, because coders discussed their coding at multiple stages to reach a consensus, the data they generated are not entirely independent. I therefore report here on my examination of potential rater effects.

To determine the extent to which coders differed systematically in applying the fidelity coding scheme, I conducted an analysis of variance for each FOI variable by coder. I limited these analyses to only students whose video data had not been double-coded for reliability purposes.

Only one omnibus F-test indicated a significant difference in fidelity data by coder—that of the variable *majerr* (whether tutors committed major errors on the initial assessment). The results of Tukey's wholly significant difference (WSD) post-hoc comparison of means suggested that coding of this variable by both coders two and four differed significantly from that of coders one and five. However, the Tukey-Kramer method for pairwise comparisons, which is preferred in cases such as this where group sizes differ, revealed no differences at the $p = .05$ level. Nonetheless, as I describe below, I controlled for the possible differences in coders' application of this particular code in analyses pertaining to the initial assessment,.

Research questions and models

In this final subsection pertaining to my analyses, I describe the models I employed to answer my research questions and thereby address three overarching goals: 1) assess MR's potential for successful scale-up (i.e., deteremine whether it was successfully implemented with fidelity), 2) test the program theory of MR, and 3) identify ways in which the MR program can be improved. After describing the findings of the main effects analysis to establish benchmarks, I first assessed whether, in general, the intervention was delivered with fidelity. Next, focusing on the first step of MR's theorized causal chain depicted in Figure 4 above, I examined the extent to which tutors' fidelity to the prescribed use of the initial assessment enabled them to correctly assign students' Learning Framework profiles at the outset of tutoring. According to the MR model, beginning in the 'right place' should lead to greater gains by maximizing the amount of instruction that matches the student's zone of proximal development (as defined by MR). Next, I employed a series of models linking FOI indices to student outcomes in order to examine the extent to which those outcomes were influenced by (a) structural aspects of MR; (b) process aspects of MR; and (c) non-MR aspects of tutoring (i.e., positive infidelity). Finally, I examined the MR program theory with respect to the relationship of increased strategies for number and improvement in mathematics achievement by testing whether improvement on the MR initial assessment (a measure of the development of new strategies) mediated the effects of treatment on more distal outcomes (i.e., Woodcock-Johnson and MR Proximal assessments).

In Chapter 4, I begin by describing the results of the main effects analysis (Smith, Cobb, Farran, Cordray, et al., 2010) to provide a foundation for interpreting findings regarding the extent to which FOI indices are related to outcomes. In these analyses, the authors accounted for the clustering of students within tutors by using a two-level hierarchical linear model (HLM). This also allowed them to control for several student characteristics often associated with variation in early mathematics achievement, including pretest score, gender, limited English proficiency, free or reduced price lunch status, and age (at the time of pretest), as well as study characteristics, including the site where students received tutoring (and where the tutors were trained) and year students received tutoring. The pretest measure was calculated as the first principal components of participants' scores on the MR proximal and each of the three Woodcock Johnson III subscales at the start of 1st grade. The model was:

$$Outcome_{ij}=\beta_{0j}+ \beta_{1j}(TREATMENT)_{ij}+ \beta_{2j}(PRETEST)_{ij} + \beta_{3j}(FEMALE)_{ij} + \beta_{4j}(LEP)_{ij} + \beta_{3j}(FRPL)_{ij} + \beta_{4j}(AGE)_{ij} + \beta_{5j}(SITE1)_{j} + \beta_{6j}(SITE3)_{j} + \beta_{6j}(YEAR2)_{ij} + u_j+ r_{ij}$$

where student i is nested within tutor j. The coefficient on *treatment*, $\beta_{1j,}$ can be interpreted as the average treatment effect. To calculate this as an effect size, the authors divided β_{1j} by the standard deviation of the outcome being analyzed, which I include in my report of the findings in Chapter 4.

In order to conserve degrees of freedom in my analyses, I did not include student characteristics that were included in the main effects analysis other than pretest scores. Nor did I include the year the student received treatment. As shown in Table 8,

characteristics were balanced across treatment conditions. Only the mean age at time of pretest was significantly different between groups; but this equated to a difference in average age of only 17 days. Additionally, student characteristics and year of treatment were not strongly correlated with study outcomes. Table 8 shows correlations between characteristics of students whose video data was selected for the fidelity analysis and the four outcome measures.

Table 8
Student characteristics by treatment condition and correlations with study outcomes

	Mean and SD by treatment condition		Correlations between outcomes and characteristics of students selected for fidelity analysis			
Characteristic	Treatment (n = 345)	Control (n = 451)	MR initial assessment (n = 102)	MR Proximal (n = 106)	WJIII-mf (n = 106)	WJIII-mr (n = 106)
sex (1 = female)	0.55 (0.50)	0.55 (0.50)	0.03	0.13	0.17	0.17
lep	0.16 (0.36)	0.12 (0.33)	0.07	0.03	–0.00	–0.06
low ses	0.64 (0.48)	0.64 (0.48)	–0.16	–0.07	0.10	–0.27*
age[a]	78.40 (4.08)	77.82* (3.95)	–0.10	–0.16	0.08	–0.16
year2	–	–	–0.06	0.01	–0.09	0.08

Note. WJIII = Woodcock Johnson III assessment of Math Fluency (mf) and Math Reasoning (mr), a combination of scores on the Quantitative Concepts and Applied Problems subtests.
* $p < .05$; [a]n = 448 for control

I did, however, include the variables in the model above pertaining to site of treatment. Site 1 included students who were tutored in the ten suburban schools, Site 2 included students tutored in the five urban schools, and Site 3 included those tutored in the five rural schools. These three categories allowed me to account for three sources of systematic variation. The first pertains to potential differences associated with types of school districts (urban, suburban, and rural). The second concerns differences at the time of the study reported by cooperating district leaders and teachers with respect to the

mathematics curricula used in each of the three sites. Differences in the extent to which the curricula used in classrooms in the different districts aligned (or did not align) with the principles of the MR program could have resulted in differential impacts on student learning. If, in one district, students' mathematics classroom experiences were similar to what they experienced in MR tutoring, students in that district could potentially have had additional opportunities to extend what they were learning in tutoring that students in other districts might not have had. Site 3 schools were employing traditional mathematics curricula at the time of the study. Both Sites 1 and 2 had relatively long histories with reform-oriented mathematics programs. However, both districts adopted a new elementary mathematics program during the second year of the study; interestingly, each district switched to what the other was previously using.

The third possible source of systematic variation stems from potential differences in training (or tutor response to training) between the two training locations (which, as explained above, corresponded to the states where the schools were located). Site 3 (rural schools) tutors received training from the same U.S. Math Recovery Council (USMRC) trainers, but separately from tutors in Sites 1 and 2, who were trained together. The research team did not directly assess the fidelity with which tutor training was delivered because it was overseen by the USMRC itself and was therefore presumably implemented with high fidelity. However, as a check on 'uptake' of training, tutors did complete the MR Tutor Knowledge Assessment (TKA) immediately following training. This paper-and-pencil, scenario-based, multiple-choice instrument was designed for the evaluation study by MR developers and members of the research team to assess a tutor's

understanding of the MR Frameworks and how to apply them in tutoring and assessment sessions.

There was a significant difference between training sites in TKA scores at the conclusion of the initial training (Green, Smith, & Neergaard, 2010). As indicated in Table 9, the average TKA score of the five tutors from the rural districts (training site A) was significantly lower than that of the thirteen tutors from the urban and suburban districts (training site B) before tutoring began. Although there was no significant difference between the two sites at the end of the two-year study, the initial difference could be interpreted as an indication of a lack of fidelity of training implementation. Again, to account for this potential source of systematic variation, my models included dummy variables representing the three sites.

Table 9
TKA scores by training site

Mean TKA score (*Total correct out of 39 questions*)	Training Site		
	Site A (n=5)	Site B (n=13)	All Tutors (n=18)
Beginning of year 1**	25.4 (sd=1.14)	30.9 (3.04)	29.4 (3.65)
End of year 1*	26.4 (2.88)	31.5 (2.63)	30.1 (3.53)
End of year 2	29.8 (3.96)	31.6 (2.53)	31.1 (2.99)

**Difference between sites statistically significant (p=.001)
* Difference between sites statistically significant (p=.02)

Question 1: Did MR tutors implement the program with fidelity? This question corresponds to my first over-arching goal, assessing a program's potential for successful scale-up. As noted above, the larger evaluation study did not find positive, lasting effects of the MR intervention, indicating that the program should not be scaled-up at this time. What remained to be answered was whether the intervention that was evaluated was,

indeed, that which the MR model prescribes. Addressing this question was important in order to determine *why* it should not be scaled-up by completing O'Donnell's (2008) test for scale-up describe in Chapter 2. To answer this question, I calculated descriptive statistics of each of the variables (as reported above in Table 6) and interpreted those statistics in terms of overall fidelity of implementation (e.g., generally high or low). To do so, I used the results of the coding of expert tutors described above as comparative benchmarks.

Question 2: To what extent is greater fidelity of implementation of MR's initial assessment associated with correct assignment of students' Learning Framework profiles at the outset of tutoring? To answer this question, I employed the following two-level hierarchical model (a), using STATA's xtixed command (to account for the clustering of students within tutors):

$$PROFPERC_{ij}=\beta_{0j} + \beta_{1j}(SITE1)_{ij} + \beta_{2j}(SITE3)_{ij} + \beta_{3j}(MAJERR)_{ij} + \beta_{4j}(MINERR)_{ij} + \beta_{5j}(INFOTOTAL)_{ij} + \beta_{6j}(CODER2)_{ij} + \beta_{7j}(CODER4)_{ij} + u_j + r_{ij}$$

where

- PROFPERC is the percentage of aspects of the MR Learning Framework on which tutor *j* correctly assigned a profile for student *i* (limited to the three aspects for which codes were reliably applied);
- SITE1 is a dummy variable coded 1 if the student was tutored in study site 1;
- SITE3 is a dummy variable coded 1 if the student was tutored in study site 3;

- MAJERR represents the total number (out of 12) of major errors committed by the tutor when administering the initial MR assessment to the student;
- MINERR represents the total number (out of 12) of minor errors committed by the tutor when administering the initial MR assessment to the student;
- INFOTOTAL represents the total (out of 5) number of aspects of the MR Learning Framework for which tutor used the initial assessment to generate sufficient information to assign a profile stage/level for the student;
- CODER2 is a dummy variable coded 1 if the student's assessment was coded by coder 2; and
- CODER4 is a dummy variable coded 1 if the student's assessment was coded by coder 4.

To answer the remaining questions, which required linking the FOI variables to student outcomes, I employed two-level multiple regression models to examine the extent to which variation in both structure- and process-oriented aspects (Mowbray et al., 2003) of FOI explains the effects of the MR intervention among those students who received treatment. In order to determine the best model for doing so, I conducted the FOI analyses in multiple layers. To address my second over-arching goal of testing MR's program theory, I first determined the impact of structural fidelity criteria, including number of lessons received, average length of a tutoring session, etc. Then, I examined the impact of process-oriented fidelity criteria, including quality of tutors' practices (i.e., 'quality of delivery') and participant responsiveness. To address my third over-arching goal of program improvement, I determined the extent to which local adaptations of the MR model (i.e., positive infidelity indices) explained variation in outcomes, separate

from other aspects of FOI (O'Donnell, 2008). In each case, I employed two models. In the first, I included all indicators separately; in the second, I substituted aggregated versions for some variables. For each pair of models, I then determined whether the model deviance significantly differed between the elaborated (all indicators included separately) and simplified (aggregates substituted) versions in order to determine the most parsimonious model for linking aspects of FOI to student outcomes.

Question 3: To what extent is greater fidelity of implementation of MR's structural aspects associated with greater student outcomes? To answer this question, I employed the following two-level hierarchical model (1a):

$$Outcome_{ij}=\beta_{0j}+ \beta_{1j}(PRETEST)_{ij}+ \beta_{2j}(SITE1)_j + \beta_{3j}(SITE3)_j + \beta_{4j}(LSSN\_NO)_{ij} + \beta_{5j}(MEANLESSTIME)_{ij} + \beta_{6j}(AVGSEALTIME)_{ij} + u_j+ r_{ij}$$

where

- $Outcome_{ij}$ is the achievement score (on the MR initial assessment, MR Proximal, WJ-III Math Fluency subtest, and WJ-III Math Reasoning subscale, a combination of scores on the Applied Problems and Quantitative Concepts subscales) at the end of first grade of student *i* in tutor *j*;
- β_{0j} is the mean achievement of tutor group/tutor *j* adjusted for student pretest, demographics, site, year and structural aspects of FOI;
- r_{ij} is a random student effect—the deviation of the student's score from the tutor's mean, assumed to be normally distributed with mean of 0 and variance

of σ^2;

- u_j is a random tutor effect—the deviation of tutor *j*'s score from the study mean, assumed to be normally distributed with mean of 0 and variance of σ^2;
- SITE1 is a dummy variable coded 1 if the student was tutored in study site 1;
- SITE3 is a dummy variable coded 1 if the student was tutored in study site 3;
- PRETEST represents the student's score on the dependent variable assessment at the beginning of first grade;
- LSSN_NO represents the total number of MR tutoring lessons the student received;
- MEANLESSTIME represents the average length of tutoring sessions the student received; and
- AVGSEALTIME represents the average number of minutes per tutoring session spent on strategy-focused tasks; and

As in the main effects analysis (Smith et al., 2010), the first variable, *pretest*, was calculated as the first principal components of participants' scores on the MR proximal and each of the three Woodcock Johnson III subscales (Math Fluency, Applied Problems and Quantitative Concepts) at the start of 1st grade. The last three variables, *lssn_no*, *meanlesstime*, and *avgsealtime*, represent the structural aspects of MR discussed above.

I followed this analysis with a similar model (1b), substituting for the three structure indicators a single aggregate, *time*. It represents the ratio of time spent on strategy-based activities to total lesson time, weighted by the number of lessons. I calculated it as follows: *time* = (*avgsealtime/meanlesstime*)**lssn_no*. Then, as explained

above, I determined whether the increase in model deviance was sufficiently small to warrant using the aggregated version of the variables in the final model.

Question 4: To what extent is greater fidelity of implementation of MR's process aspects associated with greater student outcomes? To answer this question, I created a second model (2a):

$$Outcome_{ij}=\beta_{0j}+ \beta_{1j}(PRETEST)_{ij}+ \beta_{2j}(SITE1)_j + \beta_{3j}(SITE3)_j + \beta_{4j}(ADJ)_{ij} + \beta_{5j}(TCGP)_{ij} + \beta_{6j}(BEHAV)_{ij} + \beta_{7j}(WAIT)_{ij} + \beta_{8j}(DEMO)_{ij} + \beta_{9j}(CHECK)_{ij} + \beta_{10j}(SOLIC)_{ij} + u_j+ r_{ij}$$

where

- ADJ represents tutor *j*'s average rate of adjusting the level of difficulty of tasks that, in their original form, were too difficult for student *ij*;
- TCGP represents the tutor's average rate of posing genuine problems (as opposed to tasks that were too easy or too difficult) to the student;
- WAIT represents the tutor's average rate of provided sufficient post-task wait time to the student;
- BEHAV represents the tutor's average rate of refraining from eliciting particular behaviors from the student;
- DEMO represents the tutor's average rate of refraining from demonstrating to the student methods for solving particular types of problems;
- CHECK represents the tutor's imbalance between correct and incorrect student responses in asking the student to check her or his answer, weighted by the frequency with which the tutor employed the move with the student; and

- SOLIC represents the tutor's imbalance between correct and incorrect student responses in soliciting the student's strategy, weighted by the frequency with which the tutor employed the move with the student.

I followed this analysis with a similar model (2b), substituting the nature of instruction scale described above, *noi*, for its three constituents, *wait*, *behav*, and *demo*, and, again, examined the change in model deviance.

Question 5: To what extent is higher frequency of non-MR aspects of tutoring (i.e., positive infidelity) associated with greater student outcomes? To answer this question, I created a third model (3a):

$$Outcome_{ij}=\beta_{0j}+ \beta_{1j}(PRETEST)_{ij}+ \beta_{2j}(SITE1)_j + \beta_{3j}(SITE3)_j + \beta_{4j}(REVOICE)_{ij} + \beta_{5j}(COMPARE)_{ij} + \beta_{6j}(DIFF)_{ij} + u_j+ r_{ij},$$

where

- REVOICE represents the tutor's average rate of revoicing the student's strategies;
- COMPARE represents the tutor's average rate of asking the student to examine the mathematical similarities and differences among two or more strategies; and
- DIFF represents the tutor's average rate of asking the student to solve a problem in a different way.

As before, I followed this analysis with a similar model (3b), substituting the positive infidelity scale described above, *posinf*, for its three factors, *revoice*, *compare*, and *diff*, and, again, examined the change in model deviance.

Question 6: To what extent are all aspects of fidelity of implementation combined

associated with greater student outcomes? I concluded this series of analyses by creating a final model (4) that included all of those FOI variables identified in models 1-3 as being significantly related to one or more study outcomes. I describe this model and report its results in Chapter 4.

Question 7: To what extent does student responsiveness, as measured by gains on the MR initial assessment, mediate the effect of tutoring on external mathematics assessments? This final question pertains to a particular component of MR's change model, depicted in Figure 4 above, that students' development of new and more sophisticated strategies for solving number problems will enable them to do well in regular mathematics classrooms (or, in terms of our evaluation, perform equally to their peers on more global assessments of mathematics knowledge). MR's initial assessment is the program's most direct measure of students' development of new, more sophisticated strategies in number. The main evaluation analysis described above found a significant difference in scores on the MR initial assessment at the end of first grade between students who received tutoring and those who did not. In the analysis I describe below, I address the question of whether this difference was attributable to the indirect effect of treatment by way of increasing students' strategies in number (as theorized by the MR model), or attributable to more direct effects of tutoring.

To answer this question, I conducted a mediation analysis, using the entire evaluation data set, including students in both treatment and control conditions, to determine the indirect effects (those mediated by an increase in number strategies) of tutoring on the MR Proximal and the Woodcock-Johnson Math Fluency and Math Reasoning measures. Figure 8 represents the steps of the mediation analysis, where the

arrow labeled (a) represents the effect of tutoring on gains on the MR initial assessment; (b) represents the effect of gains on the MR initial assessment on study outcomes (i.e., the WJ-III and MR Proximal assessments); (c) represents the direct effect of tutoring on study outcomes; and the product of (a) and (b) represents the indirect effect of tutoring on study outcomes, mediated by gains on the MR initial assessment.

I first created a variable representing student gains on the MR initial assessment, *assmtgain*, by subtracting the score at the beginning of the year in which the student received tutoring from the score at the conclusion of that year. Then, employing a two-

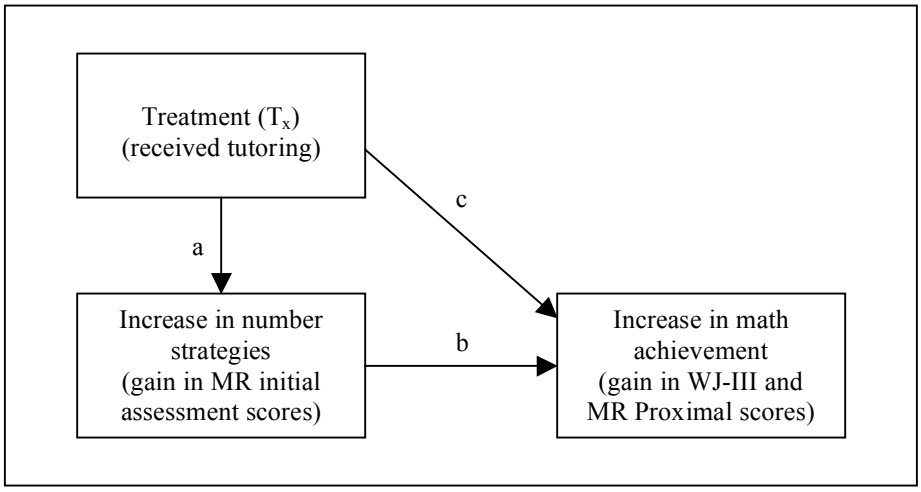


Figure 8. Mediation model.

level hierarchical model and controlling for the same variables as in the models described above, I regressed the *assmtgain* variable on *treatment*, a dummy variable indicating whether a student received tutoring or was part of the control group. The coefficient of the *treatment* variable represented the effect of tutoring on gains in MR initial assessment scores, depicted as arrow (a) in Figure 8 above. The equation was:

$$ASSMTGAIN_{ij}=\beta_{0j}+\beta_{1j}(TREATMENT)_{ij}+\beta_{2j}(PRETEST)_j+\beta_{3j}(SITE1)_j+\beta_{4j}(SITE3)_{ij}+u_j+r_{ij}$$

where

- ASSMTGAIN represents the student's change in scores on the MR initial assessment from beginning of the year in which the student received tutoring to the end of that year;
- TREATMENT is a dummy variable indicating whether the student received tutoring or was part of the control group;
- PRETEST represents the students aggregate score on the MR proximal and each of the three WJ-III subscales at the beginning of first grade as describe above (this variable did not include the student's score on the MR initial assessment); and
- SITE1 and SITE3 are, as before, dummy variables coded 1 if the student was tutored in study site 1 or study site 3, respectively.

Then, I employed a reduced form equation, regressing each of the WJ-III and MR Proximal outcomes on both *treatment* and *assmtgain*, again controlling for site and pretest:

$$Outcome_{ij}=\beta_{0j}+\beta_{1j}(TREATMENT)_{ij}+\beta_{2j}(ASSMTGAIN)_j+\beta_{3j}(PRETEST)_j+\beta_{4j}(SITE1)_{ij}+\beta_{4j}(SITE3)_{ij}+u_j+r_{ij}.$$

The regression coefficient of *treatment* represented the direct effect of tutoring on outcomes, arrow (c) in Figure 8 above, and the coefficient of *assmtgain* represented the effect of increased number strategies on outcomes, arrow (b). I then calculated the indirect effect of tutoring on outcomes by way of increased number strategies by calculating the product of the coefficient on *assmtgain*, arrow (b), and the coefficient on *treatment* from the previous model, arrow (a). To the extent that the effect of tutoring identified in the main evaluation analysis was diminished in the last equation, I could interpret this as evidence of the role that changes in number strategies play in mediating the effect of tutoring on math achievement, and thus evidence of the validity of this component of the MR change model.

I performed this analysis in two ways, 1) including all students in the study, and 2) including only those students whose MR initial assessment scores at the beginning of first grade were below the median of the study population. I performed the latter analysis in order to examine the effects of developing of more sophisticated strategies in number (as measured by the MR initial assessment) for those students who had the most room for growth.

CHAPTER 4

RESULTS

In this chapter, I present the results of my analyses by addressing each of the research questions I raised in Chapter 3. I have organized the sections according to my over-arching goals: 1) assessing MR's potential for successful scale-up (i.e., determining whether it was successfully implemented with fidelity), 2) testing the program theory of MR, and 3) identifying ways in which the MR program can be improved.

First, however, I describe the results of the evaluation study's main effects analysis (Smith, Cobb, Farran, Cordray, et al., 2010) to provide a foundation for interpreting findings regarding the extent to which FOI indices are related to outcomes. Controlling for pretest score, gender, limited English proficiency, free or reduced price lunch status, and age (at the time of pretest), as well as study characteristics, including the site where students received tutoring (and where the tutors were trained) and year students received tutoring, these authors determine the effect of tutoring on the student outcomes described in Chapter 3 at the end of the year in which students were tutored. Table 10 shows the results of this analysis. To interpret the effect of tutoring as an effect size, the authors divided the regression coefficient for treatment by the standard deviation of the each dependent variable.

The evaluation study found significant effects of Math Recovery on all end-of-year outcomes. The effect was, not surprisingly, largest (0.85) on the MR initial assessment, a tutor-administered assessment that closely resembles the content and nature

of MR tutoring. A modest effect (0.26) was found for the WJ-III Math Reasoning measure, a combination of the Applied Problems and Quantitative Concepts subscales, and for the MR Proximal. The latter result is somewhat surprising given that the MR Proximal measure was developed in consultation with the program developers to measure what students would likely learn in the course of the MR intervention. Last, the evaluation study found a small effect (0.14) of MR on the WJ-III Math Fluency measure.

Table 10
Results of main effects analysis

Assessment	Study SD of outcome	Effect size	N
MR Initial Assessment	4.00	0.85***	759
MR Proximal	2.67	0.26***	775
WJIII Math Fluency	3.56	0.14*	775
WJIII Math Reasoning (Applied Problems + Quantitative Concepts)	11.04	0.26***	775

* $p < 0.05$; *** $p < 0.001$ (two tailed tests)

Where appropriate in the results reported below, I include the above effect sizes as benchmarks in order to compare the effects of aspects of FOI on student outcomes. Specifically, I provide 'partial effect sizes' by dividing fidelity regression coefficients by the effect size determined by the evaluation analysis described above for each outcome.

Assessing MR's potential for successful scale-up

Question 1: Did MR tutors implement the program with fidelity?

To answer this question, I calculated descriptive statistics of each of the variables (as reported in Table 4 of Chapter 3) and interpreted those statistics in terms of overall fidelity of implementation (e.g., generally high or low). Additionally, where appropriate,

I compared these findings to similar results generated by applying my coding schemes to expert MR tutors' work with students. I should note that generating the results I report here required video data to which the fidelity coding schemes could be applied. The fact that I had access to video-recordings of 97% of the assessment and instructional sessions conducted by the tutors (with a majority of the missing 3% of sessions due to technical failures) indicates that the tutors adhered to the MR expectation that all sessions are video-recorded. (However, I do not have any evidence as to whether tutors used these video-recordings in accordance with MR expectations to diagnose students' current strategies and plan for subsequent lessons.)

Figure 8 shows the distributions of the variables pertaining to administration of the initial assessment, which was conducted with all 107 students (100%) who received tutoring. Each variable has been re-scaled so that a value of 1 represents perfect fidelity, and a value of 0 represents complete infidelity. The first two plots show tutors' rates of committing major and minor errors on the 12 (of 19) sections of the initial assessment for which codes were applied reliably. Means across all 107 students whose assessments were coded were 0.93 and 0.91, respectively, which can be interpreted as the percentage of the 12 sections on which those types of errors were *not* committed.

The next plot, *infototal*, represents tutors' effectiveness during the initial assessment in generating sufficient information to be able to assign a profile stage or level for students on the five (of six) aspects of the MR Learning Framework that the initial assessment covers. The mean was 0.84, indicating that, on average, sufficient information was generated for about four of the five aspects.

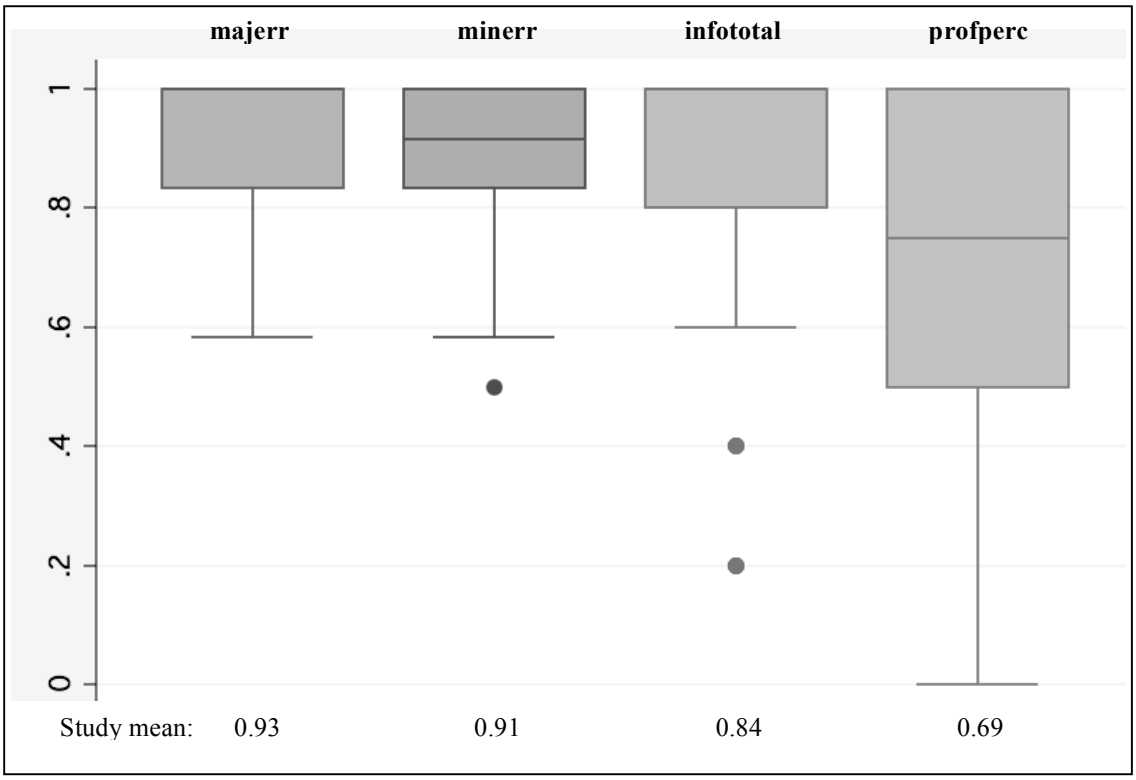


Figure 9. Distributions of initial assessment FOI variables (1 = perfect fidelity)

The fourth plot in Figure 9, *profperc*, shows tutors' effectiveness in assigning the correct stage or level on the MR Learning Framework based on the initial assessment. This variable includes three aspects of the Learning Framework that were reliably coded: SEAL, for which tutors' rate of correctly assigning the stage was 0.67; Numeral Identification, for which tutors' rate of correctly assigning a level was 0.64; and Structuring Number, for which the rate was 0.77. The mean for the aggregate *profperc* variable, depicted in Figure 9 above, was 0.69. As explained in Chapter 3, SEAL was weighted twice the other two aspects included in the *profperc* variable because of its prominence in the MR Learning and Instructional Frameworks.

The remaining FOI variables pertain to instructional sessions. Table 11 reports tutors' rates of fidelity to exposure and duration guidelines set forth by the MR model

and, in the case of number of tutoring sessions, the operational expectation (based on the average number of lessons provided to MR students in school districts other than those that participated in the evaluation study, supplied by the USMRC). The study mean of only one variable, the average length of tutoring lessons (*meanlesstime*), met the prescribed level of the MR model, but, as was the case with all three exposure and duration variables, the average length of tutoring sessions varied widely. The study means for both number of lessons (*lssn_no*) and average amount of time per lesson spent on strategy-based activities (*avgsealtime*) were significantly less than what are recommended by the program model.

Table 11
Tutor fidelity to exposure/duration

FOI variable	MR model	Operational expectation	Mean	SD	Min	Max	% meeting model expectation	% meeting operational expectation
Number of tutoring sessions (*lssn_no*)[a]	*48-60*	*42*	32.49	7.75	3	52	3.48%	11.88%
Average length of tutoring sessions in min. (*meanlesstime*)[b]	*25-30*	*(No data available)*	25.04	3	16.92	31.38	58.88%	n/a
Average time spent on strategy-based activities in min. per session (*avgsealtime*)[b]	*10-13.5 (40-45% of session)*	*(No data available)*	6.54 (26% of session)	3.30	0.7	30.79	13.08%	n/a

[a] $n = 345$; [b] $n = 107$

Figure 10 shows distributions of tutors' fidelity to the process-oriented aspects of MR, with the 15 expert tutors' mean scores on the indicators superimposed with thin black bars. The first two plots pertain to essential and unique aspects of MR: tutors' frequency of posing "genuine problems" as tasks (*tcgp*, originally intended as part of a

'zone of proximal development' scale) and frequency of adjusting the difficulty of tasks that were initially too difficult (*adj*, originally intended as part of an 'ongoing assessment' scale). On average, 62 percent of the tasks the tutors posed were genuine problems. Expert tutors posed genuine problems only slightly more frequently (65 percent). On average, study tutors adjusted task difficulty when the original tasks were too difficult at a rate of 0.83, which is considerably more frequently than the expert tutors' average rate of 0.39. This difference should not necessarily be interpreted as a lack of fidelity as adjusting task difficulty through additional scaffolding or modifying the task is consistent with the MR model. One plausible explanation for the difference between study and expert tutors' rates of adjusting task difficulty is that expert tutors were more likely to choose a third option prescribed by the MR model for responding to situations in which tasks are too difficult: directly releasing the student from the task and moving on to a new task. If this were the case, it would suggest that study tutors were more likely than experts to persist with the initial task by adjusting its level of difficulty than to abandon the task for a new, less difficult one.

The next four plots represent tutors' enactments of the 'nature of instruction' variable (*noi*) and its three constituents, frequency of providing sufficient post-task wait-time (*wait*), refraining from eliciting particular behaviors (*behav*), and refraining from directly demonstrating a method for solving a type of task (*demo*)—essential but not unique aspects of MR. As indicated by Figure 10, study tutors were slightly less likely to adhere to MR expectations for all three aspects. T-tests revealed that the differences were statistically significant for *behav* ($p < .05$) and *demo* ($p < .001$), with study tutors refraining 86 percent of the time on average for both and expert tutors refraining 93 and

98 percent of the time, respectively. On the composite *noi* variable, the difference between study and expert tutors was not statistically significant.

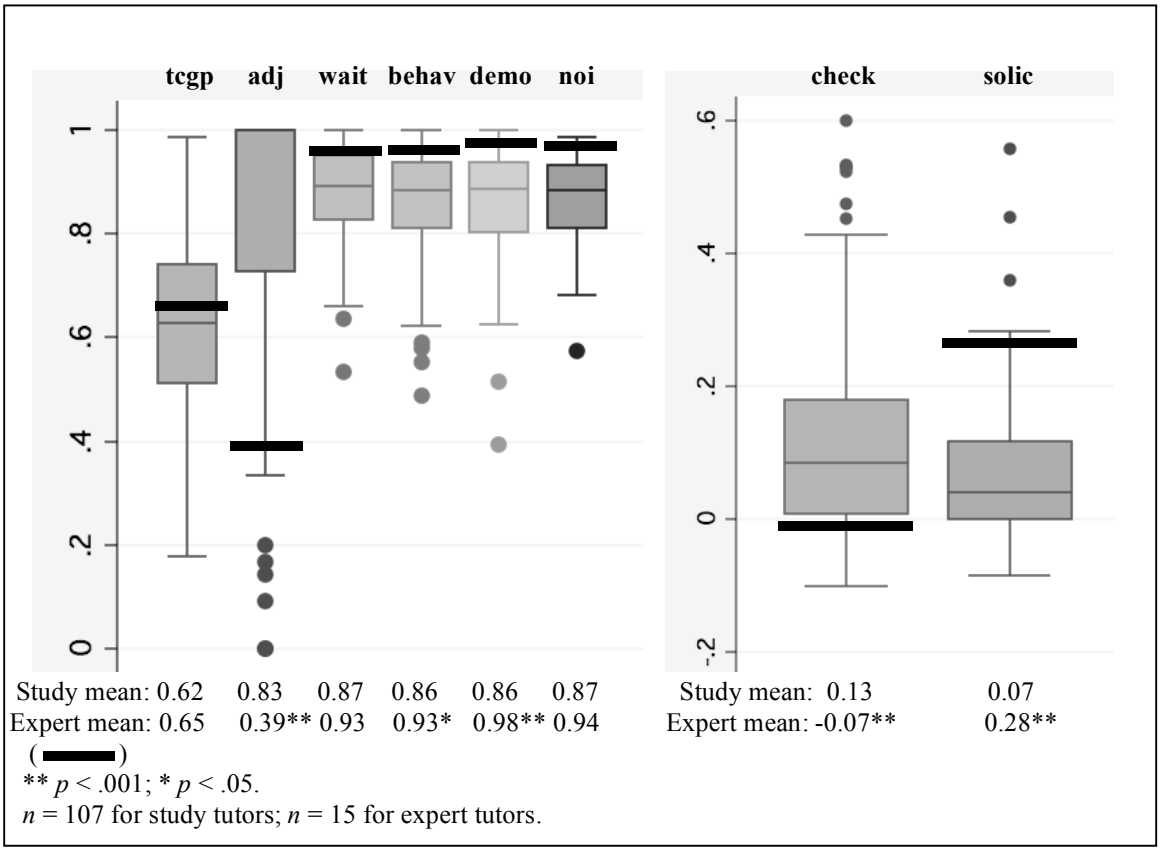


Figure 10. Distributions of MR process FOI variables with expert means

The last two plots in Figure 10, child-checking (*check*) and soliciting a strategy (*solic*) report indicators of the other essential but not unique aspects of MR. As I explained in Chapter 3, I had originally intended to include these two indicators in the nature of instruction variable. For these graphs I left them in their original form, where (1) represents asking the student to check his or her strategy or soliciting the student's strategy after a student's correct answer and (-1) indicates such tutor actions after an

incorrect answer. Therefore, as explained in Chapter 3, a mean of (0) indicates a balance between the two, which is the expectation of MR (so as not to indicate to students whether answers are correct but to require them think through their strategies and answers themselves). For both indicators there was a significant difference ($p < .001$) between study and expert tutors. However, whereas expert tutors more often balance their requests for student to check their work across correct and incorrect student responses, study tutors were more likely to strike such a balance with respect to soliciting students' strategies. Specifically, on average, experts tutors were 1.78 times as likely ([0.36+0.28]/0.36) to solicit students' strategies after correct responses than after incorrect responses. This suggests that with respect to balancing solicitations of strategies across correct and incorrect student responses, study tutors' practices were actually more aligned with the MR model than those of expert tutors.

Finally, the four plots of Figure 11 depict tutor's frequencies of practices implicitly prohibited by the MR model: revoicing students' strategies (*revoice*), asking students to compare different strategies for solving a problem (*compare*), and asking students to solve the same task in a different way (*diff*), as well as the aggregate of these three measures of positive infidelity (*posinf*). Expert tutors employed only one of the three aspects of positive infidelity with any frequency, revoicing students' strategies, and they did so more often, on average, than study tutors. However, this difference is not statistically significant. That expert tutors employed neither of the other two moves, asking students to compare strategies or solve a task in a different way, should not necessarily be interpreted as an indication that expert tutors never employ such moves. Given the limited number of expert tutor videos that were coded, and the infrequency

with which study tutors used these moves (*revoice* occurred in 21% of lessons; *compare* in 4%; and *diff* in 10%), it is not highly improbable that these are the results of chance.

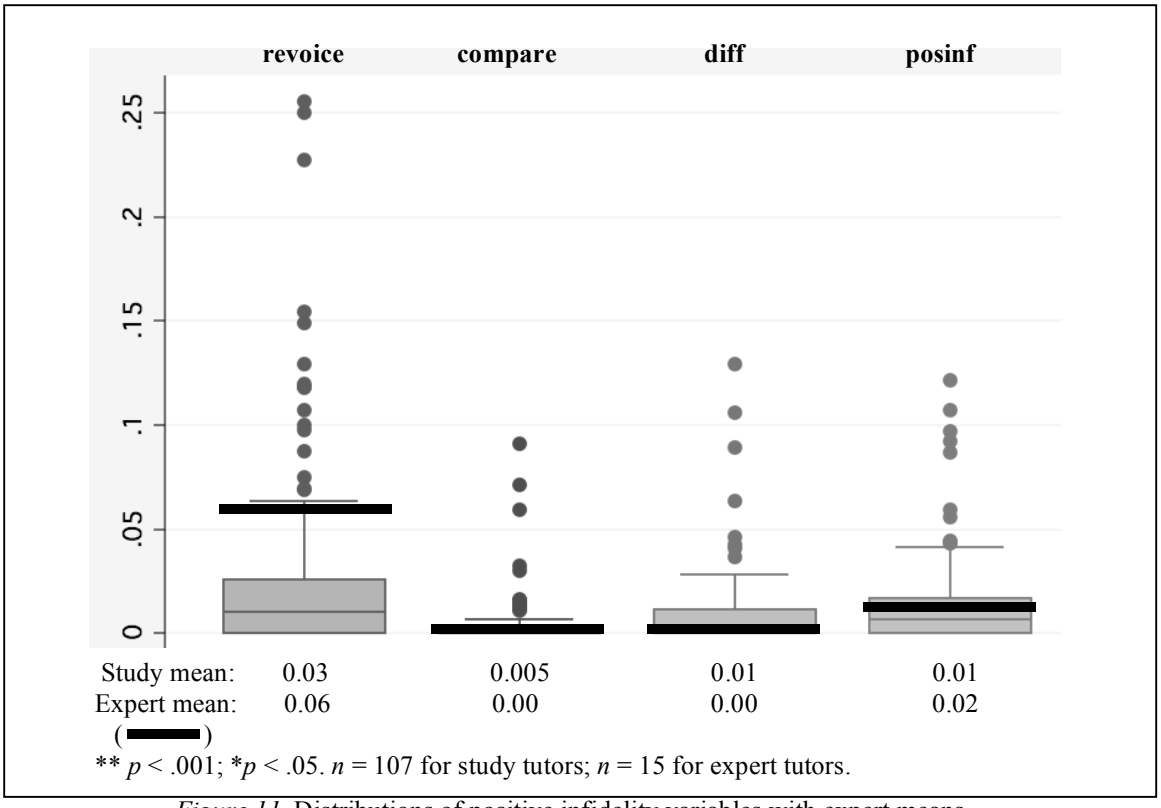


Figure 11. Distributions of positive infidelity variables with expert means

Summary

Examining the fidelity results descriptively suggests that, in general, the intervention was implemented with inconsistent fidelity. With respect to structural aspects of the intervention, the initial assessment was administered to all students at the outset of tutoring; tutors committed relatively few errors when administering the initial (scripted) assessment; and, on average, the length of lessons met MR expectations. However, FOI was questionable with respect to other structural aspects. On average, tutors' diagnoses of students' profiles on the MR Learning Framework at the outset of

tutoring were correct only about two thirds of the time; tutors spent less time per lesson on strategy-based activities than recommended by the MR model; and students received fewer lessons than recommended by the MR model.

The design of the evaluation study likely contributed to a minor extent to the last of these deviations from the model, the average number of lessons. Given the need to administer external assessments at regular intervals and to work within school calendars, the evaluation team was forced to schedule tutoring cycles such that cycles 2 and 3 in some schools afforded 1-2 days for tutoring fewer than the operational expectation described above of 42. Specifically, the schedule of assessments allowed for at least 58 days of tutoring in all sites in year 1, cycle 1, and at least 40 days of tutoring (two days shy of the operational expectation of 42) in every site for year 1, cycles 2 and 3. In year two, the cycle lengths were more balanced, allowing for at least 44 days of tutoring in all sites for all 3 cycles. Although for most cycles a sufficient number of days (at least 42) were available for tutoring, absences and school-related interruptions are, of course, unavoidable. Therefore, it would not be surprising if, for some students, the number of tutoring sessions received were somewhat smaller than the number of days available for tutoring. However, the significant difference between the operational expectation of 42 lessons and the study mean of 32.49 lessons cannot be attributed to the design constraint identified above. Instead, the difference suggests that other aspects of the implementation of MR contributed to the shortcoming in number of lessons received. I address the severity of this shortcoming below, when I report on the results of the models I described in Chapter 3, in which I controlled for number of lessons.

With respect to process aspects of MR, tutors were, for the most part, relatively successful in delivering MR lessons when compared with expert tutors. Nearly two-thirds of the tasks tutors posed were 'genuine problems'; the 'nature of tutors' instruction' (providing sufficient wait-time, refraining from behavior eliciting and demonstrating moves) accorded with MR expectations nearly 90% of the time; and tutors' choices of when to ask students to check their work or explain their strategies were only slightly skewed toward correct responses. Tutors did occasionally engage in 'positive infidelity' moves, but these were relatively infrequent (study tutors' frequency of 'revoicing' was actually less than that of expert tutors) and, as I described in Chapter 3, should, theoretically, enhance rather than detract from the impact of MR tutoring. Overall, the inconsistency in tutors' fidelity to the MR model warrants an examination of the impact of FOI on study outcomes, beginning with the initial steps in the MR theory of change.

Testing MR's Program Theory

In this section, I report on my analyses for testing the soundness of MR's program theory using the models I described in Chapter 3. I begin by reporting on the relationship between tutors' FOI of the initial assessments and the success with which they assigned students to stages and levels of the MR Learning Framework. Then, I report on the relationship between study outcomes and tutors' fidelity to both structural aspects of MR and process aspects of MR. At the end of the chapter, I return to the MR program theory to report on the relationship between treatment, student responsiveness to MR tutoring (as measured by gains on the MR initial assessment), and study outcomes.

Question 2: To what extent is greater fidelity of implementation of MR's initial assessment associated with correct assignment of students' Learning Framework profiles at the outset of tutoring? To answer this question, I employed a two-level hierarchical model (to account for the clustering of students within tutors) and regressed the percentage of aspects of the MR Learning Framework on which students' initial profiles were correctly assigned by tutors (*profperc*) on 1) rates of both major and minor errors (*majerr*, *minerr*) and 2) the total (out of 5) number of aspects of the MR Learning Framework for which the tutor generated sufficient information to assign a profile stage/level for the student (*infototal*). Additionally, I controlled for site.

Table 12 lists the results of this model. None of the assessment FOI indicators predicted the accuracy with which the tutors assigned students' initial profiles. The non-significance of the coefficients of these three variables could be a result of the restriction of range problem alluded to in Chapter 3. However, the tutors administered the initial assessment with relatively high fidelity on average, but assigned correct student profiles only two thirds of the time. This indicates that that neither tutors' adherence to correct assessment administration (relatively few errors) nor tutors' success in generating sufficient information to diagnose stages and levels on the MR Learning Framework contributed to their final assessments of students' profiles. Of the controls, only the Site 3 control was marginally significant, which is perhaps related to the suspicions described in Chapter 3 that tutors at that site were less responsive to training than those in the combined training for sites 1 and 2, and were thus less prepared to use the MR assessment and Framework as intended. The non-significance of the coefficients for

coder2 and *coder4* confirm that there was no significant rater effect with respect to the coding of the initial assessments, as described in Chapter 3.

TABLE 12
Effects of FOI of MR's initial assessment on tutors' accuracy in assigning students' Learning Framework profiles at the outset of tutoring

Variable	Coefficient	s.e.
Constant	0.57	(0.21)**
Site 1	–0.08	(0.09)
Site 3	–0.16	(0.09) #
Majerr[a]	–0.02	(0.04)
Minerr[a]	0.02	(0.03)
Infototal[a]	0.05	(0.04)
Coder 2	0.02	(0.07)
Coder 4	0.02	(0.11)

n = 107 students, 18 tutors; # $p < .10$; ** $p < .01$; [a]1 = perfect fidelity; 0 = complete infidelity

In subsequent models, I did not include the FOI variables that pertained to the initial assessment[3]. This is because there is no reason to suspect that a direct link exists between tutors' errors on the initial assessment and student outcomes after weeks of tutoring.

Question 3: To what extent is greater fidelity of implementation of MR's structural aspects associated with greater student outcomes? To answer this question, I employed the following two-level hierarchical model (1a), described in Chapter 3:

$$Outcome_{ij}=\beta_{0j}+ \beta_{1j}(PRETEST)_{ij}+ \beta_{2j}(SITE1)_j + \beta_{3j}(SITE3)_j + \beta_{4j}(LSSN\_NO)_{ij} + \beta_{5j}(MEANLESSTIME)_{ij} + \beta_{6j}(AVGSEALTIME)_{ij} + u_j+ r_{ij'}$$

[3] I did investigate the extent to which tutors' effectiveness in correctly assigning students' Learning Framework profiles at the outset of tutoring (*profperc*) was associated with their effectiveness during the first three lessons after the initial assessment in posing tasks to students that were genuine problems (*tcgp*). The idea behind this analysis was that the task selection of tutors who were more successful at assigning a profile might have a higher proportion of genuine tasks. The relationship was positive (B = 0.13, approximately 0.5 SD of the dependent variable) but non-significant ($p = .23$). However, the sample size was limited. Only 50 of the 107 students selected for FOI coding engaged in strategy-based activities during the first three lessons.

regressing students' scores on each of the outcome measures at the end of first grade on the structural aspects of MR. These included the exposure and duration variables: average number of lessons received (*lssn_no*), average length of lessons (*meanlesstime*), and average number of minutes per lesson spent on strategy-based activities (*avgsealtime*). Additionally, I controlled for students' pretest scores and site of tutoring. In model (1b), I replaced the three structure indicators with *time*, the ratio of time spent on strategy-based activities to average length of lessons, weighted by the number of lessons received.

Table 13 shows the results of these analyses for each outcome measure. The table lists the partial effect sizes for each FOI variable, calculated by diving the estimated regression coefficients by the standard deviation of the outcome measure. Table 13 also includes the effect size estimates from the main effects analysis described above for comparison.

Table 13
Results of models 1a & 1b: Influence of fidelity to structural aspects of MR on student outcomes

	Outcome							
	Initial MR assessment (*n* = 101)		MR Proximal (*n* = 105)		WJIII-mf (*n* = 105)		WJIII-mr (*n* = 105)	
(Main ES)	*(0.85)*		*(0.26)*		*(0.14)*		*(0.27)*	
model	1a	1b	1a	1b	1a	1b	1a	1b
intercept	11.94**	16.43***	8.57**	6.61***	483.38***	483.10***	452.53***	464.35***
pretest	0.34***	0.35***	0.44***	0.44***	0.44***	0.46***	0.35***	0.37***
site1	0.13	0.09	–0.17	–0.13	–0.16	–0.22	0.24	0.15
site3	0.00	–0.21	–0.50	–0.37	0.64#	0.67*	–0.28	–0.46#
lssn_no	0.01		0.01		0.03**		0.02#	
meanlesstime	0.03		–0.04		–0.04		0.02	
avgsealtime	0.11***		0.05		0.08*		0.07*	
time		0.08***		0.04#		0.08**		0.06**
log likelihood	–242.46	–243.86	–241.28	–241.74	–265.48	–265.73	–359.64	–359.70
χ^2 (2)		2.78		0.93		0.50		0.12

Note. WJIII = Woodcock Johnson III assessment of Math Fluency (mf) and Math Reasoning (mr), a combination of scores on the Quantitative Concepts and Applied Problems subtests.
# $p < .10$; * $p < .05$; *** $p < .001$

Number of lessons received (*less_no*) contributed significantly to student scores on the Math Fluency portion of the Woodcock-Johnson assessment as well as the Math Reasoning composite of Applied Problems and Quantitative Concepts. However, the magnitudes are modest. Taking the WJIII-mr outcome as an example, on average, the main effect of treatment equates to an additional 13-14 lessons, or roughly three weeks of MR tutoring. The effect of average time spent on strategy-based activities (*avgsealtime*) is more striking, as the main effect sizes for the WJIII outcomes equate to a mere 2-4 additional minutes of strategy-based activities per lesson—which would still put tutors at the low end of MR's suggested 10-13.5 minutes per lesson. Of the control variables, only pretest was consistently related to student outcomes, with Site 3 marginally significantly related to the Math Fluency subtest of the WJ-III. The differences in deviances (listed as log likelihoods) of models (1a) and (1b) for each of the outcomes are negligible. I therefore used the *time* aggregate in place of the three structure variables in the final model described below.

Question 4: To what extent is greater fidelity of implementation of MR's process aspects associated with greater student outcomes? To answer this question, in model (2a) I regressed the outcomes on the following variables: the tutor's average rate of posing genuine problems (as opposed to tasks that were too easy or too difficult) to the student (*tcgp*); the average rate by which the tutor adjusted the level of difficulty of tasks that, in their original form, were too difficult (*adj*) for the student; the tutor's average rates of 1) providing sufficient post-task wait time to the student (*wait*), 2) refraining from eliciting particular behaviors from the student (*behav*), and 3) refraining from demonstrating to the student methods for solving particular types of problems (*wait*); and the weighted

imbalance factors for the tutor's use of child checking (*check*) and solicitation of strategy (*solic*). Again, I controlled for pretest and site of tutoring. In model (2b), I replaced the three nature of instruction indicators with their composite variable, *noi* (2**wait* + *behav* + *demo*), described in Chapter 3.

Table 14 shows the results of models (2a) and (2b). Imbalances in when tutors asked students to check their solutions (more often after either correct or incorrect responses), weighted by the frequencies with which they employed the move (*check*) contributed significantly to student scores on the MR initial assessment and the Math Reasoning portion of the Woodcock-Johnson assessment. The coefficients on this variable are negative, suggesting a negative relationship between such an imbalance and student achievement. Also, tutors' rates of adjusting task difficulty (*adj*) were predictive of the MR initial assessment. However, few instances of significant relationships were found for the nature of instruction indicators. Only tutors' rates of providing sufficient wait time (*wait*) were related to WJ-III Math Reasoning scores. As described in Chapter 3, all of the nature of instruction indicators are scaled from zero (complete infidelity) to one (perfect fidelity). Therefore, a reasonable interpretation of the impact of wait time on Math Reasoning is that a ten percent increase in the tutor's provision of sufficient wait time roughly equates to the effect size of 0.27 found in the main analysis. Of the control variables, only pretest was consistently related to student outcomes, with both Site variables related to WJ-III Math Reasoning scores in model (2a) (though for Site 1 the relationship is only marginally significant).

Table 14
Results of models 2a & 2b: Influence of fidelity to process aspects of MR on student outcomes

	Outcome							
	Initial MR assessment (*n* = 101)		MR Proximal (*n* = 105)		WJIII-mf (*n* = 105)		WJIII-mr (*n* = 105)	
(Main ES)	*(0.85)*		*(0.26)*		*(0.14)*		*(0.27)*	
model	2a	2b	2a	2b	2a	2b	2a	2b
intercept	13.15***	12.52***	3.09	3.25	482.81***	482.12***	456.75***	453.03***
pretest	0.41***	0.42***	0.46***	0.46***	0.49***	0.51***	0.43***	0.43***
site1	0.43	0.41	0.13	0.14	–0.07	–0.09	0.45#	0.42
site3	–0.36	–0.35	–0.45	–0.45	0.33	0.34	–0.54*	–0.54#
tcgp	0.59	0.68	0.46	0.43	–0.87	–0.78	0.45	0.54
adj	0.57#	0.61*	0.35	0.34	–0.17	–0.13	0.21	0.27
wait	1.91		0.11		2.61		2.53*	
behav	–1.07		0.75		–1.08		–1.03	
demo	–0.09		0.53		0.05		–0.68	
noi		0.85		1.35		1.72		1.10
check	–1.10*	–1.06*	–0.95	–0.97	–0.41	–0.35	–1.28*	–1.17*
solic	–0.28	–0.19	–1.01	–1.06	1.01	1.11	0.28	0.58
log likelihood	–244.52	–245.28	–240.32	–240.40	–266.18	–266.89	–356.87	–358.96
χ^2 (2)		1.52		0.16		1.41		4.18

Note. WJIII = Woodcock Johnson III assessment of Math Fluency (mf) and Math Reasoning (mr), a combination of scores on the Quantitative Concepts and Applied Problems subtests.
# $p < .10$; * $p < .05$; *** $p < .001$

Again, the differences in deviances of models (2a) and (2b) for each of the outcomes were insignificantly small, suggesting that the *noi* composite could be used in the final model. However, as discussed in Chapter 3, the *noi* scale is potentially limited by the restricted range of all three of the indicators that comprise it. Such a restriction could also explain the lack of significance of the coefficients for tutors' solicitations of student strategies (*solic*). I therefore created an additional model (2c), for which I created a new variable, *behavbysolic*, the ratio of tutors' frequency of behavior eliciting (*behav*) to the frequency with which tutors solicited students strategies. The rationale for this variable is that the learning of students might be impacted differentially if their tutor(s) had equally high rates of behavior eliciting (prohibited by the MR model), but very

different rates of soliciting students' strategies (encouraged by the MR model). By adjusting the rate of behavior eliciting by the rate of soliciting strategies, I could better estimate the extent to which tutors who frequently elicited behaviors were engaging students in so-called guessing games. That is, according to the MR model, if a tutor frequently elicits behavior but never asks the student how (s)he solved a problem, the student is likely engage in guessing games fairly quickly; but if the tutor offsets behavior eliciting by also frequently asking the student how (s)he solved tasks, then the negative impact of eliciting behaviors is potentially diminished. Across students selected for the fidelity analysis, the variable *behavbysolic* had a mean of 1.33 (ratio of frequency of eliciting behaviors to frequency of soliciting strategies) and a standard deviation of 1.44 (minimum 0, maximum 6).

Thus, model (2c), in which I retained the *wait* indicator from model (2a) because it had been significantly related to one outcome, was as follows:

$$Outcome_{ij}=\beta_{0j}+ \beta_{1j}(PRETEST)_{ij}+ \beta_{2j}(SITE1)_j + \beta_{3j}(SITE3)_j + \beta_{4j}(ADJ)_{ij} + \beta_{5j}(TCGP)_{ij} + \beta_{6j}(WAIT)_{ij} + \beta_{7j}(CHECK)_{ij} + \beta_{8j}(BEHAVBYSOLIC)_{ij} + u_j+ r_{ij}.$$

Table 15 shows the results of model (2c), compared with those of (2a) above. The new variable, *behavbysolic*, was significantly related to MR initial assessment and WJ-III Math Reasoning scores. The magnitude of the coefficient for the latter suggests that the effect size of 0.27 found in the main analysis equates to a change from soliciting student strategies equally as often as eliciting behaviors (ratio = 1) to eliciting behaviors 2.8 times as often soliciting strategies (ratio = 2.8, or roughly one SD from the mean ratio).

The *wait* indicator was no longer significantly related to any of the outcomes. However, the relationships between *check* and the MR initial assessment and the WJ-III Math Reasoning composite remained. Differences in model deviation generally favored model (2c), though, again, the differences were small. Therefore, in the final model reported below, I retained *check* and *behavbysolic*, as well as *adj* (because the results of models 2a and 2b suggested a relationship to the MR initial assessment) as the variables representing process aspects of MR.

Table 15
Results of models 2a & 2c: Influence of fidelity to process aspects of MR on student outcomes

	Outcome							
	Initial MR assessment (*n* = 101)		MR Proximal (*n* = 105)		WJIII-mf (*n* = 105)		WJIII-mr (*n* = 105)	
(Main ES)	*(0.85)*		*(0.26)*		*(0.14)*		*(0.27)*	
model	2a	2c	2a	2c	2a	2c	2a	2c
intercept	13.15***	14.18***	3.09	5.44*	482.81***	482.47***	456.75***	450.97***
pretest	0.41***	0.40***	0.46***	0.47***	0.49***	0.47***	0.43***	0.41***
site1	0.43	0.35	0.13	0.10	–0.07	–0.08	0.45#	0.39
site3	–0.36	–0.28	–0.45	–0.41	0.33	0.39	–0.54*	–0.49#
tcgp	0.59	0.46	0.46	0.24	–0.87	–0.83	0.45	0.50
adj	0.57#	0.33	0.35	0.22	–0.17	–0.20	0.21	0.24
wait	1.91	0.25	0.11	0.44	2.61	1.27	2.53*	0.64
behav	–1.07		0.75		–1.08		–1.03	
demo	–0.09		0.53		0.05		–0.68	
check	–1.10*	–1.11*	–0.95	–1.10	–0.41	–0.36	–1.28*	–1.22*
solic	–0.28		–1.01		1.01		0.28	
behavbysolic		–0.10#		–0.03		–0.10		–0.15**
Log likelihood	–244.52	–239.25	–240.32	–239.21	–266.18	–263.64	–356.87	–351.71
χ^2 (2)		–2.56		1.75		–0.18		–4.52

Note. WJIII = Woodcock Johnson III assessment of Math Fluency (mf) and Math Reasoning (mr), a combination of scores on the Quantitative Concepts and Applied Problems subtests.
# $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$

Having reported on the relationships between structure and process aspects of MR to student outcomes, I now describe the results of my analyses pertaining to potential program improvement, namely the 'positive infidelity' indicators described in Chapter 3.

Following this report, I describe the results of the final model.

Program Improvement

Question 5: To what extent is higher frequency of non-MR aspects of tutoring (i.e., positive infidelity) associated with greater student outcomes?

To answer this question, in model (3a) I regressed the outcomes on the tutor's average rates of revoicing the student's strategy (*revoice*), asking the student to examine the mathematical similarities and differences among two or more strategies (*compare*), and asking the student to solve a problem in a different way (*diff*), again controlling for pretest and site of tutoring. In model (3b), I replaced the three positive infidelity indicators with their composite variable, *posinf*, described in Chapter 3.

Table 16 shows the results of models (3a) and (3b). Of the positive infidelity indicators, the results of model (3a) suggests that asking the student to solve a problem in a different way (*diff*) is most strongly related to student outcomes. For the WJ-III Math Reasoning composite, the coefficient on *diff* (which is scaled from zero to one) suggests that the effect size of 0.27 found in the main analysis equates to just a 3.5 percent increase in the frequency with which tutors utilize such a move. Interestingly, although only marginally significant, the relationship between asking the student to compare two or more strategies (*compare*) and the MR initial assessment is negative.

The comparison of deviance of models (3a) and (3b) warrants the use of the simplified version of the positive infidelity indicators. Although the results of model (3b) show that the composite positive infidelity scale, *posinf*, was only marginally

significantly related to one outcome, WJ-III Math Reasoning, I used this variable in the final model, described in the next section.

Table 16
Results of models 3a & 3b: Influence of positive infidelity on student outcomes

	Outcome							
	Initial MR assessment (*n* = 101)		MR Proximal (*n* = 105)		WJIII-mf (*n* = 105)		WJIII-mr (*n* = 105)	
(Main ES)	*(0.85)*		*(0.26)*		*(0.14)*		*(0.27)*	
model	3a	3b	3a	3b	3a	3b	3a	3b
intercept	19.19***	19.12***	7.39***	7.47***	485.36***	485.33***	468.67***	468.77***
pretest	0.40***	0.43***	0.45***	0.47***	0.50***	0.53***	0.39***	0.40***
site1	0.20	0.26	–0.06	–0.06	–0.10	–0.05	0.22	0.23
site3	–0.54#	–0.50	–0.47	–0.50	0.40	0.42	–0.60*	–0.61*
revoice	–0.44		–1.90		–0.27		0.67	
compare	–9.47#		4.30		–6.43		2.01	
diff	5.91		10.50*		7.42		7.46*	
posinf		–1.29		2.59		0.99		6.02#
log likelihood	–248.37	–250.60	-240.82	–242.93	-268.96	–270.35	–360.91	–362.03
χ^2 (2)		4.46		4.22		2.78		2.24

Note. WJIII = Woodcock Johnson III assessment of Math Fluency (mf) and Math Reasoning (mr), a combination of scores on the Quantitative Concepts and Applied Problems subtests.
# $p < .10$; * $p < .05$; *** $p < .001$

Additional Questions

Question 6: To what extent is are all aspects of fidelity of implementation combined associated with greater student outcomes? To answer this question, I created a the final model (4) in order to simultaneously estimate the relationships between outcomes and all variables identified as significant predictors by models 1-3:

$$Outcome_{ij}=\beta_{0j}+ \beta_{1j}(PRETEST)_{ij}+ \beta_{2j}(SITE1)_j + \beta_{3j}(SITE3)_j + \beta_{4j}(TIME)_{ij} + \beta_{5j}(ADJ)_{ij} + \beta_{6j}(CHECK)_{ij} + \beta_{7j}(BEHAVBYSOLIC)_{ij}\ \beta_{8j}(POSINF)_{ij} + u_j+ r_{ij}.$$

I interpreted the coefficients on fidelity variables as the average increase in outcome, attributable to each aspect of FOI. Again, to relate these to the effect sizes identified by the main analysis described at the beginning of this chapter, I divided each coefficient by the standard deviation of the outcome being analyzed. Table 17 shows the results of model (4).

Among the three control variables, students' pretest scores were strongly related to all outcomes, Site 1 had no significant relationship with any outcome, and Site 3 (the five rural schools, whose tutors were trained separately from those in Sites 1 and 2) was significantly related to only the WJ-III Math Fluency test. The positive coefficient of the last of these suggests that, with all other variables in the model held constant, Site 3 students' Math Fluency scores were, on average, considerably higher than those of students in the other two sites.

Table 17
Results of full model (4): Influence of all aspects of FOI on student outcomes

	Initial MR assessment (*n* = 101)	MR Proximal (*n* = 105)	WJIII-mf (*n* = 105)	WJIII-mr (*n* = 105)
(*Main ES*)	*(0.85)*	*(0.26)*	*(0.14)*	*(0.27)*
intercept	16.00***	6.02***	483.59***	462.98***
pretest	0.31***	0.42***	0.42***	0.33***
site1	0.08	–0.07	–0.25	0.15
site3	–0.06	–0.27	0.76*	–0.30
time	0.08***	0.04	0.08**	0.05**
adj	0.36	0.36	0.02	0.41
check	–0.32	–0.81	–0.10	–0.79#
behavbysolic	–0.13*	–0.03	–0.12#	–0.15**
posinf	–0.75	3.93	0.06	6.10*
log likelihood	–232.93	–237.97	–261.38	–347.16

Note. WJIII = Woodcock Johnson III assessment of Math Fluency (mf) and Math Reasoning (mr), a combination of scores on the Quantitative Concepts and Applied Problems subtests.
# $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$

The composite *time* variable significantly predicted all outcomes but the MR Proximal, with the relationship to Math Reasoning being the greatest. To interpret the partial effect sizes for *time* listed in Table 17, consider two hypothetical students, both of whom received 32 lessons at an average length of 25 minutes per lesson (both study means). But for one student the average number of minutes per lesson spent on strategy-based activities was 6.5 (the study mean), while for the other student, it was 10 (the lower bound of the MR model expectation). The *time* value for the first student would be 8.32 (32 * 6.5 / 25), and for the second student, 12.8 (32 * 10 / 25). This difference of 4.48 is equivalent to an effect size of 0.22 (4.48 * 0.05) on the WJ-III Math Reasoning measure, suggesting that an increase of only 3.5 minutes per lesson spent on strategy-based activities (in this hypothetical, but highly representative case) would have nearly the same effect on Math Reasoning scores as treatment was found to have in the main evaluation analysis. In a similar scenario, the main effect size for the WJ-III Math Fluency measure is achieved by an 8-minute increase in time spent on strategy-based activities per lesson, or 14.5 minutes per lesson, just beyond the upper bound of the MR expectation.

Among the process variables, *behavbysolic*, the ratio of the tutor's frequency of behavior eliciting to that of soliciting the student's strategies, was significantly related to the MR initial assessment and Math Reasoning, and marginally significantly related to Math Fluency, with coefficients very similar to those of model (2c) above. Neither child checking, *check*, nor adjusting task difficulty, *adj*, was significantly related to any outcome, although the relationship between *check* and Math Reasoning was marginally significant. The positive infidelity scale, *posinf*, significantly predicted only Math

Reasoning. The partial effect size reported in Table 17 above (6.10) suggests that the main effect size for that outcome (0.27) is equivalent to a four percent (0.27 / 6.10) increase in positive infidelity moves. This finding provides strong evidence that the practice of MR tutoring could be improved by incorporating aspects of mathematics instruction identified in the literature of recent years such as those included as positive infidelity indicators.

Having described the results of my layered analyses of the associations between aspects of FOI and student outcomes, I now return to the MR program theory to examine one of its fundamental components more broadly. According to the MR change model (depicted in Figure 4 in Chapter 3), students' development of multiple strategies for solving number problems will enable them to do well in regular mathematics classrooms (or, in terms of our evaluation, perform equally to their peers on more global assessments of mathematics knowledge). MR's initial assessment is the program's most direct measure of students' development of new, more sophisticated strategies in number. The results of the analyses reported above indicate that aspects of FOI are related to students' performance on the initial MR assessment at the end of the year in which they received tutoring. These aspects include the fraction of lesson time spent on strategy-based activities (*time*), behavior eliciting and soliciting students' strategies (*behavbysolic*), and, although only marginally significant, the frequency with which tutors adjust the difficulty of tasks that were initial too difficult for students (*adj*). Having identified these relationships, I now step back to examine whether, as suggested by the MR program theory, students' development of new strategies in number mediates the relationship between treatment and more distal mathematics assessments.

Question 7: To what extent does student responsiveness, as measured by gains on the MR initial assessment, mediate the effect of tutoring on external mathematics assessments? As described in Chapter 3, I conducted a mediation analysis, using the entire evaluation data set, including students in both treatment and control conditions, to determine the direct and indirect effects (those mediated by an increase in number strategies) of tutoring on the MR Proximal and the Woodcock-Johnson Math Fluency and Math Reasoning measures. Controlling for the same variables as I did in the models described above, I regressed *assmtgain* (the student's gain on the MR initial assessment from the beginning of the year in which tutoring was received to the end of that year) on *treatment*, a dummy variable indicating whether a student received tutoring or was part of the control group, to determine the effect of tutoring on gains on the MR initial assessment. Then, I regressed the four outcome scores on both *treatment* and *assmtgain* to determine the direct effect of tutoring on outcomes (the regression coefficient of *treatment*) and the indirect effect of tutoring on outcomes by way of increased number strategies (the product of the coefficient on *assmtgain* and the coefficient on *treatment* from the previous model). As also explained in Chapter 3, I performed this analysis in two ways, first 1) including all of the students selected for fidelity coding, and then 2) limiting it to just those students whose MR initial assessment scores at the beginning of first grade scores were below the median of the study population.

Table 18 shows the results of the mediation analysis. The regression coefficients reported in the first row are the result of the first model, which predicted *assmtgain* based on *treatment*. They are not related to the study outcome variables, but are repeated across the columns for ease of comparison. The coefficients in the second and third rows are the

result of the reduced form equation, which predicted each of the three study outcome variables based on both *assmtgain* and *treatment*. The fourth row lists the products of the first and second, the indirect effect of tutoring on study outcomes, by way of gains on the MR initial assessment. In the fifth row are sums of the values in rows three and four, the total effect (both direct and indirect) of tutoring on study outcomes. In the last row of Table 18, I repeat the values of row five, recalculated as effect sizes (by dividing by the standard deviation of the dependent variable).

Table 18
Results of mediation analysis

	MR Proximal		WJ-III Math Fluency		WJ-III Math Reasoning	
	All (*n* = 749)	< 50% (*n* = 368)	All (*n* = 746)	< 50% (*n* = 367)	All (*n* = 746)	< 50% (*n* = 367)
Direct effect of T_x on assmtgain (a)	3.13 (0.24)***	3.32 (0.36)***	3.13 (0.24)***	3.32 (0.36)***	3.13 (0.24)***	3.32 (0.36)***
Direct effect of assmtgain on outcome (b)	0.18 (0.02)***	0.22 (0.03)***	0.25 (0.03)***	0.35 (0.05)***	0.69 (0.08)***	0.91 (0.11)***
Direct effect of T_x on outcome (c)	0.14 (0.17)	–0.04 (0.23)	–0.34 (0.25)	–0.83 (0.34)*	0.80 (0.61)	–0.23 (0.86)
Indirect effect of T_x on outome (a*b)	0.57	0.73	0.77	1.17	2.15	3.01
Total effect of T_x on outcome (a*b+c)	0.70	0.69	0.44	0.34	2.95	2.78
Total effect of T_x on outcome as effect sizes	0.26	0.26	0.12	0.10	0.27	0.25

*** $p < .0001$

The results of the mediation analysis confirm that students' development of new strategies in number mediates the effect of tutoring on study outcomes. The effects of tutoring on end-of-first-grade WJ-III and MR Proximal scores, which were identified in the main evaluation analysis, were diminished in the last equation. Gains on the MR initial assessment were the significant predictor of study outcomes; coefficients on *treatment* were not significantly different from zero. This was true for all outcomes and

in both versions of the analysis: 1) all students included, and 2) limited to only those students whose MR initial assessment scores were below the study population median at the beginning of first grade, with just one exception. Interestingly, the direct effect of tutoring on study outcomes for students who began below the study median on the MR initial assessment was significantly negative. However, the total effect, as listed in the last row of Table 18, was positive, though small.

These results also confirm the findings of the main effects analysis, described at the beginning of this chapter. The small to modest effect sizes listed in Table 18 are nearly identical to those shown above in Table 10. The advantage of the mediation analysis, however, is that it helps articulate the composition of those effects, by examining steps of the theorized causal chain underlying the MR model. The results of the analysis provide evidence of the validity of this component of the MR change model, that to the extent that MR tutoring 'works,' it is through its contribution to students' development of new strategies in number.

CHAPTER 5

DISCUSSION

In this chapter I discuss implications of the analyses and findings I have presented. I highlight key aspects of the work that were particularly helpful in assessing fidelity of implementation of Math Recovery in order to illustrate how these steps might be accomplished in fidelity studies of other unscripted interventions, and a discussion of the feasibility of assessing FOI of unscripted programs (e.g., lessons learned etc.). First, however, I begin with a discussion of five limitations of the study.

Limitations

First, as explained in Chapter 3, although the coders applied our instruments reliably with respect to most fidelity indicators, it was not possible to code a few aspects of the implementation of MR with sufficient reliably for inclusion in my analyses. Any intended variable that was not used could have potentially added to the explanatory power of my models. However, one exclusion was likely more limiting than the others: whether each teaching procedure (i.e., set of related tasks) a tutor used with a student was aligned with the student's current profile on the MR Learning Framework. This aspect of tutors' practices was covered to some extent by the *tcgp* variable, the percentage of tasks posed to a student that were genuine problems (neither too easy nor too difficult). But the procedure alignment indicator could have provided useful information regarding tutors'

success in aligning MR's two frameworks, the Learning Framework and the Instructional Framework.

Similarly, the study did not include a measure of the fidelity of tutors' ongoing updating of student profiles on the MR Learning Framework. This limitation was not the result of a low rate of coder agreement. It was instead a consequence of the decision to reduce the demands on tutors so as not to over burden them with tasks related to the evaluation study but extraneous to their work as MR tutors. For purposes of the evaluation study, the tutors made copies of all assessment and lesson video-recordings as well as all lesson plans and sent them to the evaluation team. Although most lesson plans listed a profile for the students (i.e., numbers indicating the student's stage or levels on the MR Learning Framework), it was not clear how often tutors updated these profiles and therefore the entries could not be used as reliable data. An assessment of the extent to which tutors adhered to this aspect of the MR model would have required daily submissions from all 18 tutors concerning all 60 students being tutored at any given time. Given the other evaluation study-specific requirements placed on tutors, such a request did not seem feasible. As a result, one missing component from the fidelity assessment was the frequency and accuracy with which tutors updated students' profiles on the MR Learning Framework, one representation of the quality of tutors' ongoing assessment of students' current strategies and capabilities.

Third, as stated in Chapter 3, the absence of detailed data on control students' mathematics learning opportunities was likely another limitation of the study. Ideally, I would have assessed the extent to which regular classroom instruction resembled that of MR tutoring in order to calculate the "achieved relative strength" of the intervention

(Cordray & Hulleman, 2009). That is, I would have compared the achieved strength of treatment delivered in tutoring and classrooms (as compared to the MR model) to the "treatment" received by those students not in tutoring. Such an assessment was beyond the scope of the small-scale evaluation of MR (largely because the study sites were remote), and assessing program differentiation (Dane and Schneider, 1998) in this case was likely not as crucial as in other evaluations. The evaluation compared outcomes of students who received the supplemental MR tutoring with those of students who did not receive the supplement; none of the students who remained on the wait list received MR tutoring (or any other mathematics intervention); and none of the students who received MR tutoring received any other mathematics intervention. Nonetheless, data on the nature of classroom learning opportunities afforded to students, in both treatment and control groups, and the extent to which those opportunities aligned with MR tutoring would have been helpful in explaining differences between treatment conditions, as well as between tutoring sites.

Fourth, as shown in Chapter 4, treatment condition was a significant predictor of gains on the MR initial assessment, which in turn mediated the impact of tutoring on more distal outcomes. However, aspects of how the MR initial assessment works as a meaningful measure remain unclear. First, the assessment was administered by MR tutors, rather than the external assessors who administered all other outcome measures. Therefore, the reliability of students' scores on the MR initial assessment is dependent, in part, on the accuracy with which tutors assigned the scores. Second, the psychometric properties of the MR initial assessment have not been examined. Its multiple items were designed to help tutors diagnose students' thinking with respect to six constructs (i.e.,

stage and levels of the MR Learning Framework), but the assessment's construct validity and reliability have not been formally assessed. Nor is it clear whether some components of the assessment are more predictive of gains on distal outcomes than others. Although the mediation analysis reported in Chapter 4 sheds light on the mechanisms by which treatment students made gains as a result of MR tutoring, this line of work cannot be extended until the MR initial assessment has been thoroughly examined.

Last, as described in Chapter 4, the design of the evaluation study likely contributed to a minor extent to limiting the possible number of tutoring sessions for some students in some cycles. Given the need to administer external assessments at regular intervals and to work within school calendars, the evaluation team was forced to schedule tutoring cycles such that cycles 2 and 3 in some schools afforded 1-2 days for tutoring fewer than the operational expectation described above of 42. However, as already addressed, the significant difference between the operational expectation of 42 lessons and the study mean of 32.49 lessons cannot be attributed to this design constraint. Other aspects of the implementation of MR must have contributed to the shortcoming in number of lessons received for many students.

Interpreting the Main Effects With Respect to Fidelity Findings

As described in Chapter 1, O'Donnell (2008) described the scaling-up decision-making process in terms of two measures of an initial effectiveness study: outcomes and fidelity. Based on her argument (as depicted by the 2 x 2 matrix in Figure 1), if study outcomes are low and fidelity of implementation is high, then the results are attributable

to program ineffectiveness. However, if FOI is low, then such attribution cannot be made. In either scenario, of course, the results of the MR evaluation indicate that the program should not be adopted at a wide scale at this time. As has been shown, the larger evaluation study in which the fidelity study was conducted found that the Math Recovery intervention failed to produce positive, lasting effects on student achievement. The estimated effect sizes of tutoring on mathematics achievement scores at the end of first grade were small to modest. The results of the analyses that I have reported in this dissertation help to explain the reasons.

The question is whether MR should not be scaled up at this time because 1) it does not "work," or because 2) the program in its intended form has not actually been evaluated (i.e., during the evaluation, the intervention was not implemented with sufficiently high fidelity to warrant claims regarding its effectiveness). The results I reported in Chapter 4 indicate that the answer is not clear-cut. To address this issue, I revisit the three overarching goals of the dissertation. The first two, 1) assessing MR's potential for successful scale-up (i.e., determining whether it was successfully implemented with fidelity), and 2) testing MR's underlying program theory, relate to the question posed above. The findings that relate to the first two goals have implications for the third, 3) identifying potential areas for program improvement. If outcomes were low and FOI was high, this might suggest modifying fundamental aspects of the program, whereas if both outcomes and FOI were low, this might suggest modifying the supports for implementation and then re-evaluating the program. In other words, beyond identifying reasons why the evaluation did not find large effects of the MR intervention, the answer to the above question can have implications for how MR developers and

evaluators proceed.

Assessing Math Recovery's potential for successful scale-up

A primary goal of effectiveness studies is to examine how programs operate in actual use, in the hands of practitioners who have chosen to adopt and implement them. This contrasts with efficacy trials, in which the goal is to assess the effectiveness of a program under ideal conditions, in the hands of experts—often developers of the program being evaluated (Dorland, 1994; O'Donnell, 2008). O'Donnell (2008) argued that assessments of FOI serve different purposes within these two different forms of evaluation. In efficacy trials, fidelity of implementation is monitored and controlled throughout the duration of the study in order to ensure that a high level of fidelity is maintained and to identify both the most critical program components (Mowbray et al., 2003) and components that require refinement (Dane & Scheider, 1998). In effectiveness studies, where the over-arching goal concerns generalizability, assessments of FOI serve as a measure of feasibility of large-scale implementation (Darrow, 2009; Dusenbury et al., 2003) by determining reasonable expectations for how closely to its intended form a program can be implemented in natural settings.

The evaluation of MR is a case of the latter of these two types of studies. Although its scale was small, the effectiveness study examined the impact of MR tutoring provided by newly trained, first-time tutors in districts that had not previously adopted the program. This framing of the larger evaluation study implies that the FOI findings reported in Chapter 4 should be interpreted as indications of the feasibility of implementing the MR program as intended in natural settings, and the results found in the

main analysis as a measure of what magnitude of effects can be expected from such an implementation.

As reported in Chapter 4, many aspects of the MR intervention were implemented according to model (or operational) expectations. The MR initial assessment was administered to all students, with tutors committing relatively few errors in doing so; nearly all assessment and instructional sessions were video-recorded by tutors; the average length of lessons, 25.04 minutes, met the MR guideline; tutors' average rate of posing genuine problems (neither too difficult nor too easy) to students was comparable with that of expert tutors (although this varied across students more than other fidelity indicators); tutors were successful with the majority of students in adjusting the difficulty of tasks that, when initially posed, were too difficult; and tutors' "nature of instruction" (rates of providing sufficient wait time and refraining from eliciting behaviors and demonstrating methods) was, on average, comparable with that of expert tutors' practices (with relatively little variation across students).

However, a number of aspects of the MR intervention were implemented with questionable levels of fidelity: initial diagnoses of students' profiles on the MR Learning Framework (0.67 rate of correct profile assignment); total number of lessons provided to students (32.49 study mean as opposed to the operational expectation of 42); average amount of time per lesson spent on strategy-based activities (6.54 min/lesson study mean as opposed to the model expectation of 11.75); and tutors' use of child checking (an average imbalance toward correct responses of 0.13 among study tutors as opposed to the operational expectation observed among expert tutors of only 0.07 toward incorrect responses). Additionally, study tutors' average ratio of frequency of behavior eliciting to

frequency of soliciting students' strategies was 1.33 as opposed to expert tutors' average ratio of 0.10.

Overall, the fidelity of implementation of MR was inconsistent with respect to model expectations. The minimal effect sizes estimated by the evaluation study are potentially attributable, at least in part, to this shortcoming. Also, the inconsistency in FOI could suggest that implementing MR tutoring as intended, with the current forms of tutor professional development and support, is not feasible in natural settings. However, settings differ. With respect to MR, as shown in Table 19, significant differences in FOI existed between training locations. Controlling for students' free and reduced priced lunch status, limited English proficiency and pretest scores, Site 3 (rural districts) tutors' fidelity to number of lessons and average time spent on strategy-based activities, as well as frequency of instances of positive infidelity, were significantly below those of tutors in Sites 1 and 2 (suburban and urban districts, trained together in a different location as Site 3 tutors). These differences are potentially related to a number of factors. I discuss three of the most likely candidates below.

First, differences in tutors' responsiveness to training, as discussed in Chapter 3, could have contributed to levels of FOI over all, and especially tutors' uses of the MR Frameworks and process aspects of MR tutoring (e.g., refraining from eliciting student behaviors, employing positive infidelity moves). The likelihood of such a difference in tutor responsiveness to training is increased by the fact that training was provided by the same USMRC representatives in both training locations.

Second, local school calendar limitations could have contributed to differences in numbers of lessons received. For example, for three schools in Site 3, the school years

were longer by, on average, approximately one week. However, such a difference could have amounted to, at most, two tutoring lessons more per cycle, and, although these Site 3 schools had a greater number of possible days for tutoring, students in Site 3, on average, received fewer lessons than students in Sites 1 and 2.

Table 19
Fidelity and student characteristic means by training location

FOI VARIABLE	Sites 1 and 2 combined mean and SD	Site 3 mean and SD	Sig[a]	Sig after controlling for student characteristics[b]
Number of lessons (*lssn_no*)	34.03 (6.32)	27.79 (9.65)	***	***
Average strategy-based time per lesson (*avgsealtime*)	7.64 (3.00)	3.59 (2.00)	***	***
Adjusting task difficulty (*adj*)	0.85 (0.22)	0.77 (0.31)		
Child checking (*check*)	0.15 (0.15)	0.10 (0.10)	#	
Behavior eliciting adjusted for strategy solicitation (*behavbysolic*)	1.15 (1.38)	1.84 (1.52)	*	
Positive infidelity (*posinf*)	0.02 (0.03)	0.003 (0.006)	**	**
STUDENT CHARACTERISTIC				
Free/reduced price lunch	0.60 (0.49)	0.76 (0.05)	**	
Limited English proficiency	0.16 (0.37)	0.01 (0.11)	***	
Pretest	–0.04 (0.84)	–0.31 (0.77)	*	

# $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$; [a]result of t-test (105 df); [b]result of multiple regression

Last, differences in tutors' classroom practices prior to the evaluation study could have influenced FOI to the extent that 1) prior instructional practices aligned or did not align with MR tutoring, and to the extent that 2) tutors incorporated those practices into their MR tutoring. Because no data was collected on tutors' classroom practices prior to the evaluation study, this consideration is speculative. However, as described in Chapter 3, study sites differed in types of mathematics curricula used in classrooms: whereas the schools in Sites 1 and 2 had employed reform-oriented mathematics curricula for several

years prior to the MR evaluation, schools in Site 3 had employed traditional mathematics curricula. If differences in tutors' instructional histories (and experiences with various mathematics curricula) influenced FOI, such differences likely contributed to tutors' responsiveness to training as well.

Therefore, estimates of feasibility must take into consideration the relevant types of resources that exist in settings in which programs are implemented (or are being considered for implementation). For example, given the impact that increased time spent on strategy-based activities has on the effects of the MR intervention, adopting school districts may want to consider the extent to which potential MR tutors are experienced with mathematics curricula that emphasize, in addition to other aspects of early number, supporting students in developing strategies for solving number problems. The significance difference in *avgsealtime* between training locations noted in Table 19 might have been related to differences in tutors' inclinations to engage students in such content, which, given the differences in participating districts' curricular histories described in Chapter 3 and noted above, might have been influenced by practices that tutors developed in mathematics classrooms with particular instructional materials prior to becoming MR tutors.

Testing MR's program theory

Again, given the small to modest effects found by the main evaluation, the question is whether the program was not implemented well or is not effective in its current form, with its current sources of support. Having pointed to aspects of the evaluation that possibly give credence to the former explanation, I now turn to the latter

and discuss implications of the fidelity study with respect to MR's program theory. My focus is primarily on the MR tutoring model, but, at the end of this section, I broaden my focus to consider other aspects of the program theory, including the current forms of implementation supports for MR tutors.

First, in addition to determining more nuanced estimates of implementation feasibility, the FOI assessment that I conducted helps identify what the effects of the intervention could potentially be if it is implemented as intended. To determine the impact that questionable levels of FOI might have had on the results of the evaluation of MR, I estimated what the outcome means for the treatment group would potentially have been had the program been implemented with levels of fidelity meeting 'operational expectations'—not necessarily the model expectations, but levels observed in video recordings of expert tutors or based on implementation of MR in other settings—and then compared the effect of treatment in that simulated version to the effects estimated by the main evaluation analysis. To do so, I first estimated the 'treatment-with-fidelity' mean outcome scores by multiplying the regression coefficients from my full fidelity model that were at least marginally statistically significant by the values observed as operational expectations. In order to estimate simulated effect sizes, I had to account for what would have been a change in the study population standard deviations. I therefore centered the treatment outcome scores on the mean identified in the previous step (which slightly increased the study population standard deviations). Next, I calculated the potential effect sizes by dividing the difference between the 'with-fidelity' treatment group means and control means by the combined standard deviation of the dependent variable, calculated with the new, centered means.

Table 20 shows the results of the simulation. I substituted values for the three FOI variables that were significant predictors of the MR initial assessment and the Math Fluency and Math Reasoning portions of the Woodcock-Johnson assessment (the table excludes the MR Proximal outcome, since, as reported in Chapter 4, none of the FOI variables significantly predicted those scores). For the *time* variable, I multiplied the operation expectation of 42 lessons (observed in other settings in which MR has been implemented) by the ratio of time spent on strategy-based activities (11.75 minutes per lesson, median model expectation) to the length of an MR lesson (27.5 minutes, median model expectation). This calculation yielded a value of 17.95 (as compared to the study mean of 8.36). For the *check* variable, I substituted the observed imbalance of child checking among expert tutors, 0.16 (which actually exceeds the study mean of 0.14). For the *behavbysolic* variable, I again used the mean value observed among expert tutors, 0.10 (as compared to the study mean of 1.33). Multiplying these valued by the regression coefficients estimated by the final fidelity model in Chapter 4 and adding them to the intercepts produced the potential 'treatment-with-fidelity' means listed in Table 20.

The potential effect size estimates are considerably larger than those found in the main effects analysis, which are listed at the bottom of Table 20. Although the estimates provide an indication of the potential impacts of MR tutoring if implemented with higher levels of fidelity, they should be interpreted with caution for a number of reasons. First, MR tutoring was provided with the levels of fidelity substituted in the analysis above for only one of the 107 students selected for the fidelity assessment. Therefore, the simulation assumes that the linear relationships identified in my original fidelity analyses extend beyond nearly all the observed values within the study. Second, my attempt to

adjust for the change in standard deviations in outcomes is almost certainly insufficient in accounting for the change in the data's error structures had the program been implemented with such high levels of fidelity. Third, using only those regression coefficients from the final fidelity model that were significant predictors of each outcome ignores the covariance those variables had with non-significant predictors. However, among the MR fidelity variables, correlations were not high. For only one pair, *behavbysolic* and *adj*, was the correlation above 0.2 (0.23).

Table 20
Potential effects of treatment with sufficient levels of FOI

FOI variable (study mean)	Value substituted	Source	Regression coefficient: MR initial assmt	WJ-mf	WJ-mr
intercept			16.00	483.59	462.98
time (8.36)	17.95	Composite time variable: 42 lessons (operational expectation), and 27.5 min. average lesson length with 11.75 min. focused on strategy-based activities (model expectations)	0.31	0.28	0.57
check (0.14)	0.16	Expert training video mean (operational expectation)			–8.74
behavbysolic (1.33)	0.10	Expert training video mean (operational expectation)	–0.50	–0.43	–1.62
		Potential treatment-with-fidelity mean	21.49	488.58	473.69
		Control mean	14.15	484.82	465.20
		SD of outcome, recalculated to incorporate estimated treatment means	4.00	3.72	11.14
		Potential ES estimate	1.84	1.06	0.77
		Main analysis ES estimate	0.85	0.14	0.26

With these caveats acknowledged, I assert that the estimated potential effect sizes provide at least a rough estimate of the impact that the lack of FOI might have had, and help to further explain the effects estimated by the main evaluation analysis in light of the

findings regarding FOI reported in Chapter 4. Specifically, the estimates suggest that MR potentially 'works' (i.e., has significantly larger effects on student achievement—at least those at the end of the year in which tutoring was provided—than those found by the evaluation study) if it is implemented as intended. The fidelity assessment also affords a closer look at some of MR's core components and the contribution they make in the theorized causal chain.

As reported in Chapter 4, the results of my analyses validated one key aspect of the MR change model: the effects of tutoring on mathematics achievement are mediated by students' development of new, more sophisticated number strategies. But other aspects of the MR program theory were not similarly corroborated. For example, in spite of their generally correct administration of the initial assessment, tutors assigned the correct MR Learning Framework profile at the outset of tutoring only two-thirds of the time on average. This suggests that correct use of the initial assessment is much less related to adherence to the protocol than it is to tutors' individual capabilities in using the MR Learning Framework to make accurate judgments. Also, tutors' rates of posing genuine problems did not significantly predict any student outcomes, including gains on the MR initial assessment. This indicates that the changes in students' number strategies, which are related to gains in mathematics achievement, are not the results of high or low percentages of task time spent at the "cutting edge" of students' thinking. Instead, changes in students' scores on the MR initial assessment are most strongly related to the average number of minutes spent on strategy-based activities per lesson. Thus, it matters more that students have opportunities to engage in strategy-based activities than does the rate with which "genuine problems" are posed.

Considering the MR program theory more broadly, the results of the fidelity assessment also suggest examining the assumption of the MR model that tutors improve their practice over time—largely the result of repeated use of the Frameworks, time with students, and reflection. The lack of significance of the Year 2 dummy variable in my analysis suggests otherwise; no significant changes in FOI were found over the course of the two-year study. As described in Chapter 3, MR expects that new tutors in a district will meet as a cohort for two hours each month during the school year. The MR leader responsible for their training typically attends three of these meetings during the year and also conducts individual coaching sessions with tutors during these site visits. The results of my analyses suggest that the influence of such professional development opportunities on tutors' practices (i.e., fidelity to MR tutoring expectations) may be limited.

Taken together, the results of my analyses suggest that the soundness of the components of the MR program theory, including its tutoring model, change model and current forms of tutor implementation support, is inconsistent. My estimates of potential, 'with-fidelity' effect sizes suggest that the effects of MR tutoring might have been larger with higher levels of FOI. And the results of my mediation analysis suggest that the mediational components of the MR change model are valid. However, other components of the MR program theory, including both aspects of tutors' practices with students (e.g., the influence of "genuine problems") and broader characteristics of the program (e.g., ongoing tutor professional development) do not appear to contribute to the overall MR model as assumed by program developers.

Identifying potential areas for program improvement

As stated above, the evaluation of MR was an effectiveness study. Therefore, the FOI assessment serves as a measure of feasibility (Darrow, 2009; Dusenbury et al., 2003) by determining reasonable expectations for how closely to its intended form the MR program can be implemented in natural settings. However, as stated in Chapter 1, MR is a relatively new program. Its developers are still very much involved in program promotion and implementation, and the evaluation team has considerable access to those developers at a time in the program's history before it has been implemented on a large scale, or institutionalized to an irreversible degree. The fidelity assessment I conducted can therefore point to potential areas of program improvement, or critical components that deserve more attention with respect to implementation. In this section, I discuss three such areas.

First, in order to identify critical implementation components that require attention, and to determine the feasibility of boosting levels of FOI, we must consider the impact of the various components in relation to each other. Figure 12 lists the results of a analysis similar to the one summarized in Table 20 above, with the *time* variable disaggregated. Holding the values listed above for the other FOI variables constant, I estimated the impact of varying levels of average number of lessons (*lssn_no*) and average number of minutes per lesson spent on strategy-based activities (*avgsealtime*) separately—both of which were significant predictors in model (1a) in Chapter 4. I examined four combinations of levels of these two variables. For *lssn_no*, I used the operational expectation of 42 lessons and the study mean of 32.49. For *avgsealtime*, I used the median model expectation of 11.75 and the study mean of 6.54.

Number of lessons (*lssn_no*)		Min/lesson spent on strategy-based activities (*avgsealtime*)			Potential ES estimate				
OE (42)	SM (32.49)	ME (11.75)	SM (6.54)		MR initial assmt		WJ-mf		WJ-mr
	●		●		1.11		0.22		0.01
	●	●			1.67		0.59		0.28
●			●		1.29		0.53		0.24
●		●			1.85		0.89		0.51
Main analysis ES estimate					0.85		0.14		0.26

Figure 12. Potential effects of treatment with varying levels of FOI of structural aspects of MR (OE=operational expectation; ME=model expectation; SM=study mean)

The results shown in Figure 12 suggest that fidelity to model expectations for amount of time spent on strategy-based activities alone, with the number of lessons held at the study mean, would have had a significant impact on the effects of tutoring (and more so than the opposite combination). Given that these indicators represent structural (as opposed to process) aspects of the program, they could likely be improved in future implementations of MR with more explicit expectations and guidelines. In other words, the feasibility of implementing MR with greater fidelity with respect to these structural aspects is high, and the estimates listed in Figure 12 suggest that the benefits would be considerable.

It is possible that the impact of time spent on strategy-based activities is not merely a matter of allocating time for such activities, but also depends on how the remaining lesson time is coordinated with the strategy-based activities. During the training provided by the USMRC to our fidelity assessment team, the trainers emphasized that all of a tutor's instructional choices should be motivated by designing lessons that will support students in increasing the sophistication of their number strategies (i.e.,

increase their Stage of Early Arithmetical Learning, or SEAL, on the MR Learning Framework). Based on our interpretation of the training, tutors should devote lesson time to other aspects of the Learning Framework only in service of this primary goal. For example, if a student is asked to complete the task of adding 12 and 3, and responds with "twelve, ferteen, fifteen, sixteen," such a response suggests that the student is developing (or has developed) a 'counting-on' strategy for addition, but needs extra support in this particular area of the Forward Number Word Sequence (FNWS). The student's mispronunciation of "thirteen" is causing the student to skip "fourteen" in the sequence. In this case, the problem is not the student's strategy for solving the problem, but simply the words to use when using the counting-on strategy. Therefore, the tutor should devote lesson time to supporting the student in solidifying this range of the FNWS.

In the above illustration, the tutor discovered the need to work on the FNWS with the student by engaging the student in a strategy-based activity. In a majority of the video data selected for the fidelity assessment, this was not the approach that tutors took. Instead, many tutors repeatedly began lessons with activities that focused on non-strategy-based aspects of the MR Learning Framework, as though they were practicing 'the basics' first before working up to strategy-based tasks. In these cases, it seemed that tutors viewed the aspects of the MR Frameworks as discrete components, seemingly working from an incremental model of how working on these components would impact students' learning. Therefore, although time spent on strategy-based activities is a structural aspect of MR fidelity, in future implementations of the program, more explicit expectations regarding this aspect should include a more explicit conceptual rationale as well.

The second area of the MR program that the FOI assessment indicates could be improved concerns tutors' uses of the MR initial assessment and how they link its results to assigning students' profiles on the MR Learning Framework. As stated above, tutors committed relatively few errors when conducting initial assessments and, in general, generated sufficient information about students' thinking with respect to the relevant aspects of the MR Learning Framework. However, their assignments of students' profiles were not sufficiently accurate. In short, adhering to the assessment protocol does not guarantee accurate results. This finding suggests that both the initial assessment and training in using it should be re-examined. MR developers could invest time and resources in increasing the reliability of the initial assessment, which would require increasing the extent to which it can be applied systematically to produce results that are more easily mapped onto operational definitions of stages and levels of the MR Learning Framework. More likely, however, developers will need to provide tutors with more (ongoing) training in using the MR Frameworks.

As described in Chapter 3, the models of children's learning in number are grounded in research on early number learning. Therefore, it is not surprising that two weeks of training is not sufficient in enabling tutors to master their application. But, given that the Frameworks lie at the heart of the MR program, implementing the intervention with greater fidelity will require considerable attention in supporting tutors in using these tools effectively. Results of the TKA assessment, described in Chapter 3, suggest that tutors' understanding of the Frameworks does not significantly improve as a result of more tutoring; tutors will likely need more direct forms of support beyond the

initial training, and the limited follow-up opportunities currently provided by the program.

The third area for potential program improvement concerns the instances of 'positive infidelity' in tutoring practices that were incorporated into the fidelity assessment: forms of practice that align with recent research on mathematics teaching but are (at least implicitly) prohibited by the MR model. The results reported in Chapter 4 suggest that employing such moves as revoicing, asking students to solve tasks in different ways, and asking students to compare strategies had a significant impact on students' WJ-III Math Reasoning scores, which are a composite of the Quantitative Concepts and Applied Problems WJ-III subtests. To determine the potential impact that incorporating such practices into the MR model could have, I conducted a similar analysis as that reported described above. With values for *adj*, *check*, and *behavbysolic* held at the mean values of expert tutors observed in training video recordings, I estimated the impact of varying levels of positive infidelity (*posinf*) on WJ-III Math Reasoning scores. Additionally, I included varying levels of the *time* variable, examining eight combinations of values of the two variables. For *posinf*, I used the study minimum (0.00), the operational expectation of 0.02 (the mean observed in expert training videos), the 95th percentile of study tutors (0.06), and the maximum value observed among study tutors (0.12). For *time*, I used the operational expectation of 17.54 and the study mean of 8.36.

Table 21 shows the results of this analysis. The estimates suggest that, even within the study mean for *time*, the impact of employing positive infidelity moves in just six percent of tasks posed (the rate at the 95th percentile of study tutors) could have a greater impact on Math Reasoning scores than the effect of tutoring determined in the

main analysis. If the operational expectation concerning strategy-based time is met, the potential impact of employing positive infidelity moves is quite large.

Table 21
Potential ES estimates for WJ-III Math Reasoning by rates of positive infidelity and time

Rate of positive infidelity (*posinf*)		Ratio of strategy-based time to length of lesson, weighted by number of lessons (*time*)	
Source	Value substituted	Study mean (8.36)	Operational expectation (17.54)
Study minimum	0.00	-0.03	0.46
Expert training video mean (operational expectation)	0.02	0.09	0.58
Study 95th percentile	0.06	0.33	0.82
Study maximum	0.12	0.70	1.19
Main analysis ES estimate		0.26	0.26

In short, the findings suggest that MR tutoring could be improved by incorporating high-quality instructional practices identified in recent research on mathematics classroom teaching and learning. Doing so would require redefining the role of the MR tutor, from one who chooses and poses tasks to one who plays a more proactive role in assisting students in developing new and more sophisticated arithmetical strategies. Although this would represent a significant shift in how MR tutors are expected to support students' learning, it does not require a radical departure from the theories of learning underlying the MR program. Indeed, the work from which the positive infidelity moves were drawn is founded on similar theories and commitments concerning students' learning in mathematics: that students need opportunities to engage in sense-making, to build on their current understanding, and to make connections among mathematical ideas (e.g., Carpenter & Lehrer, 1999; Gravemeijer, 1994; 2004; NCTM, 2000; Putnam, Lampert, & Peterson, 1990; Sfard, 2003). The results of the fidelity

assessment do not suggest that incorporating positive infidelity moves must be at the expense of other aspects of MR tutoring. Indeed, higher rates of positive infidelity were positively (although not strongly) correlated with other aspects of FOI. Therefore, the improvements I propose should be viewed as a refinement, rather than a restructuring.

Conducting Fidelity Studies of Unscripted Interventions

In addition to the three over-arching goals of the dissertation addressed in the previous section, the fidelity study that I have described is a case that can be used to examine the feasibility of conducting fidelity studies of unscripted interventions in general—an endeavor that has not been addressed in the literature. In this section, I discuss aspects of my work in relation to three facets of fidelity of implementation described in Chapter 2: 1) FOI criteria, 2) adaptation, and 3) general guidelines for conducting FOI assessments.

Criteria for assessing fidelity

As described in Chapter 2, researchers consistently name the five criteria for assessing fidelity proposed by Dane and Schneider (1998): exposure, adherence, quality of delivery, participant responsiveness, and program differentiation. With the exception of program differentiation, which, as argued in Chapter 3, was not as applicable to the evaluation of MR as it is to other interventions, attending to each of these criteria was helpful in developing a framework for assessing FOI of MR. Additionally, Mowbray and colleagues' (2003) structure-process distinction, which O'Donnell (2008) applied to the

five criteria above, was helpful in both linking the FOI assessment to the MR program theory and situating the assessment of MR fidelity within the larger field represented by the FOI literature.

The fidelity criteria that related to structural aspects of MR included exposure/duration and adherence. With respect to the first, exposure/duration, I found that without specified time allotments (which are often provided in scripted interventions), time spent on different types of activities can vary widely by program provider. Therefore, it is important to not only measure exposure and duration by *type* of activity, but also to clarify at the outset whether the program model specifies any such time allotments. With respect to the second, I found that adherence might involve judgments on the part of program providers. To reliably assess the accuracy of such judgments requires that coders be thoroughly trained in the program itself. The training the MR fidelity coders received in the program was undoubtedly helpful; and yet, even with training, it was still not possible to assess a few potentially important indicators reliably and they had to be dropped from the analyses (e.g., alignment of tutors' choices of teaching procedures and students' profiles, as discussed above).

The fidelity criteria that related to process aspects of MR included participant responsiveness and quality of delivery. With respect to the first, I found that identifying and articulating the MR program theory and change model helped in identifying the most important aspects of student responsiveness to treatment. A proximal outcome, the MR initial assessment, also served as a measure of an important step in the theorized MR causal chain. Other interventions that are similarly diagnostic in nature could potentially have 'built-in' assessments of student responsiveness to treatment as well. With respect

to the second process criterion, I found O'Donnell's (2008) re-conception of 'quality of delivery' to be critical. As described in Chapter 2, Dane and Schneider (1998) described quality of delivery as a characteristic of program providers that potentially moderates the impact of FOI. But O'Donnell argued that, in evaluations of education programs, quality of delivery refers to whether the teacher delivers the program in the ways intended by developers, and should therefore be *included in* (rather than moderating the impact of) FOI indices. In the case of MR, and likely many other unscripted, cognitively-based interventions, the program developers consider *how* the program is delivered to be just as important as (or perhaps inextricably linked to) *what* is delivered. Assessing how a program is delivered, and the extent to which the delivery meets model expectations, requires the development of instruments that can be used to assess the quality of important dimensions of teachers' instructional practices.

Adaptation

In coding for instances of "positive infidelity" (Cordray & Hulleman, 2009), I identified specific practices that, if included in the MR model, might increase student outcomes. But gauging how large the impact of incorporating these practices into the model might potentially be (rather than merely marking their presence in tutors' practices) required identifying positive infidelity candidates a priori and building them into the MR fidelity coding instruments. These instances of positive infidelity are a form of local adaptation (Blakely et al., 1987), but one that, at least implicitly, contradicts the current program theory of the intervention. Therefore, their inclusion was not induced through analyses of previous implementations of MR, but was driven by recent research

on mathematics teaching. The usefulness of this suggests that in assessing FOI of unscripted interventions, where adaptation is often encouraged, coding schemes should be extended to include forms of practice identified in the literature as high quality that are not included in the model of the program being evaluated.

Methods for assessing fidelity

In this last section, I revisit some of the guidelines (Nelson et al., 2010; O'Donnell, 2008) I followed in conducting the MR fidelity study, and discuss them in terms of assessing FOI of an unscripted intervention. In doing so, I highlight aspects of the work that I perceive to have been particularly helpful or challenging.

With respect to the first guideline describe by these authors, identifying the program theory and core components, the evaluation team's close work with program developers in identifying core intervention components was necessary and proved to be extremely helpful. In retrospect, it would have been even more helpful to also incorporate practitioners' configural models of fidelity at an early stage in instrument development. For example, had I interviewed the 12 MR expert video raters early in the study, any aspects of tutors' practices that they consistently highlighted as exemplary or as clear contradictions of MR (i.e., 'red flags') could have been included in my coding schemes. For the second guideline, operationally defining program constructs and variables, in addition to working with program developers, it was helpful to operationalize core components at two levels, by identifying both model expectations (what would be achieved under ideal circumstances) and 'operational expectations' (what has actually been achieved in natural settings). For the latter, communicating with the USMRC and

obtaining video recordings of expert tutors was helpful in interpreting the levels of fidelity achieved by study tutors.

With respect to developing instruments, a third guideline I followed in my fidelity study, 'in-house' instruments can orient the initial stages of instrument development. In the case of MR, I encountered at least two instruments for assessing tutors' practices (for training or coaching purposes) on which I could draw. Although these were not official, USMRC-endorsed forms, they provided insights into the operationalizations and schemes that expert practitioners used to assess enactments of the program. Input of coders was also helpful in finalizing coding instruments. Collaborating with the coding team to further refine operationalizations and coding decisions helped strike a balance in our instruments between a thorough accounting of important program components and feasibility of use. As alluded to above, training fidelity coders in the intervention itself before training them to use the coding schemes and coding process was an important step in the fidelity study. Given the complexity of the intervention, and thus the complexity of the coding scheme, coders' understanding of MR's underlying theory and rationale often helped in assessing the appropriateness of tutors' judgments.

We made an important choice with respect to a fourth guideline, sampling, in response to an impasse our coding team encountered. As explained in Chapter 3, although the coders had begun to apply codes in a fairly consistent manner, reaching agreement on the units of data for applying the codes proved to be an insurmountable challenge. Therefore, we theoretically sampling from the (already randomly sampled) fidelity data by focusing solely on portions of lessons in which tutors engaged students in strategy-based activities. Agreement on code-able chunks of data were reached more

easily for these excerpts, which (a) represented the most important aspects of the intervention's change model, and (b) provided opportunities to employ the full coding scheme. Challenges such as this are likely to arise in fidelity assessments conducted within evaluations of unscripted interventions, since the programs themselves may provide no natural units of analysis. This was particularly true with respect to MR, where we applied some codes at the teaching procedure level and others at the finer grained level of tasks.

Conclusion

At the conclusion of any program evaluation, researchers must answer two questions: 1) how well was the program implemented? and 2) how effective was the program? As demonstrated in the previous sections, from a FOI perspective—and particularly from the program theory perspective I have taken in this dissertation—these questions become, 1) what level of FOI (in comparison to model expectations) is feasible? and 2) how effective might the program have been? The answers to both of these questions point to the third goal of the dissertation: identifying potential areas for program improvement.

Many potentially high-quality interventions are unscripted, require considerable tailoring by implementers, and rely on teacher knowledge and professional development. As we work to rigorously evaluate such programs, we need to develop reliable fidelity measures that are both feasible to use and true to program components. This will assist evaluators in adequately linking measures of fidelity of implementation to outcomes in

order to more accurately determine the relative strength of interventions (Cordray & Pion, 2006) and provide feedback to developers that will help in improving programs' effectiveness (Dusenbury et al., 2003). I hope that this dissertation clarifies a number of critical aspects of this work that inform others who attempt to evaluate and assess the FOI of complex interventions.

APPENDIX A

MR VALIDITY SCORE SHEET

You have been randomly assigned the following video clips to view and rate.

1.1 & 2.1 Assessments:

- NJ
- CO
- QW
- UB
- IX
- LE
- BN
- XR

Full Lessons:

- TS
- MT
- DF
- AQ
- PD
- WI
- GK
- KM

SEAL/Tens & Ones excerpts:

- SU
- EP
- VC
- JL
- RA
- HG
- OV
- FH

By writing in the pseudonyms with which the videos are labeled, please rank, from highest to lowest (1 being the highest), the tutors' enactments of Math Recovery as intended, where (1) is the highest. Additionally, for each video, please indicate with an (X) the **one** category you would use to describe that assessment or instructional lesson: *excellent*, *good*, *fair* or *poor*.

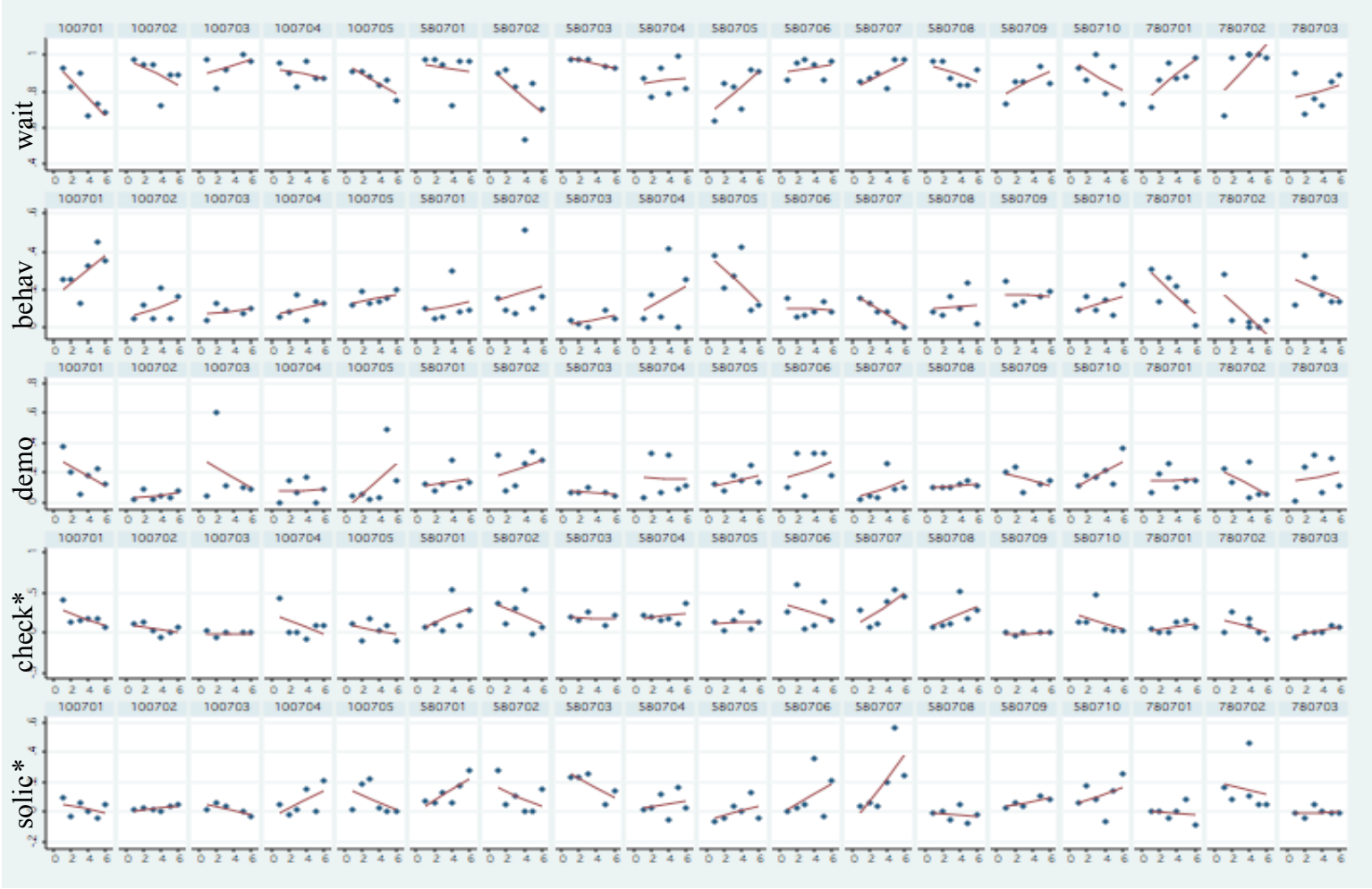
1.1 & 2.1 Assessments					**Full Instructional Lessons**					**SEAL/T&O Excerpts**				
	Category					Category					Category			
Rank	Excellent	Good	Fair	Poor	Rank	Excellent	Good	Fair	Poor	Rank	Excellent	Good	Fair	Poor
1)					1)					1)				
2)					2)					2)				
3)					3)					3)				
4)					4)					4)				
5)					5)					5)				
6)					6)					6)				

Please return this form either electronically to c.munter@vanderbilt.edu, or by mail to:

Chuck Munter
[redacted].
Nashville, TN 37209

APPENDIX B

AGGREGATED NATURE OF INSTRUCTION INDICATORS BY TUTOR OVER THE SIX CYCLES OF THE TWO YEARS



*Not included in the final NOI variable

REFERENCES

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

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

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

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

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

Bickman, L. (1987). The functions of program theory. *New Directions for Program Evaluation, 33*, 5-18.

Blakely, C., Mayer, J. P., Gottschalk, R. G., Schmitt, N., & Davidson, W. S. (1987). The fidelity-adaptation debate: Implications for the implementation of public sector social program. *American Journal of Community Psychology, 15*, 253-269.

Botvin, G. J., Dusenbury, L., Baker, E., James-Ortiz, S., & Kerner, J. (1989). A skills training approach to smoking prevention among Hispanic youth. *Journal of Behavioral Medicine, 12*, 279-296.

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

Carpenter, T. P., & Lehrer, R. (1999). Teaching and learning mathematics with understanding. In E. Fennema & T. A. Romberg (Eds.), *Mathematics classrooms that promote understanding* (pp. 19-32). Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

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

Connell, J.P. & Kubisch, A.C. (1999) Applying a theory of change approach to the evaluation of comprehensive community initiatives: progress, prospects and problems, in: K. Fulbright-Anderson, A.C. Kubisch & J.P. Connell (Eds) *New Approaches to Evaluating Community Initiatives. Volume 2: Theory, measurement and analysis* (Queenstown, The Aspen Institute).

Cordray, D. S. & Hulleman, C. (2009, June). Assessing intervention fidelity: Models, methods and modes of analysis. Presentation at the Institute for Education Sciences 2009 Research Conference, Washington, D.C.

Cordray, D. S. & Pion, G. M. (1993). Psychosocial rehabilitation assessment: A broader perspective. In R. L. Glueckauf, L. B. Sechrest, G. R. Bond, & E. C. McDonel (Eds.), *Improving Assessment in Rehabilitation and Health.* Newbury Park, CA: Sage Publications.

Cordray, D. S. & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. R. Bootzin & P. E., McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation (pp 103-124).* Washington, DG: American Psychological Association.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*(1), 23-45.

Darrow, C. (2009). Measuring fidelity in preschool interventions: A microanalysis of fidelity instruments used in curriculum interventions. Vanderbilt University.

Dorland, W. A. (1994). *Dorland's illustrated medical dictionary* (28th ed.). Philadelphia: W. B. Saunders.

Duncan, G. J., Claessens, A., & Engel, M. (2004). *The contribution of hard skills and socio-emotional behavior to school readiness*. Retrieved August 12, 2009, from http://www.northwestern.edu/ipr/people/duncanpapers.html.

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

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). Implementation research: A synthesis of the literature. Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network (FMHI Publication #231).

Franke, M. L., Kazemi, E., & Battey, D. (2007) Mathematics teaching and classroom practice. In F. K. Lester, (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 225-256). Reston, VA: NCTM.

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

Fuson, K. C. (1990). Issues in place-value and multidigit addition and subtraction learning and teaching. *Journal for Research in Mathematics Education, 21*, 273-280.

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

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

Gravemeijer, K. (1994). Educational development and developmental research in mathematics education. *Journal for Research in Mathematics Education, 25*, 443-471.

Gravemeijer, K. (2004). Local instruction theories as means of support for teachers in reform mathematics education. *Mathematical Thinking and Learning, 6,* 105-128.

Gresham, F. M. (1989). Assessment of treatment integrity in school consultation and prereferral intervention. *School Psychology Review, 18*, 37-50.

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

Hulleman, C. S. & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength." *Journal of Research on Educational Effectiveness, 2*, 88-110.

Lipsey, M. W. (1993). Theory as method: Small theories of treatments. *New Directions in Program Evaluation, 57*, 5-38.

Loucks, S. F. (1983, April). *Defining Fidelity: A cross-study analysis.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Mills, S. C. & Ragan, T. J. (2000). A tool for analyzing implementation fidelity of an integrated learning system. *Educational Technology Research and Development, 4*, 21-41.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*, 315-340.

National Research Council. (1995). *National Science Education Standards.* Washington, DC: National Research Council.

National Council for Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2010, March). A procedure for assessing fidelity of implementation in experiments testing educational interventions. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness, Washington, D.C.

O'Connor, M. C. & Michaels, S. (1993). Aligning academic task and participation status through revoicing: Analysis of a classroom discourse strategy. *Anthropology and Education Quarterly, 24,* 318-335.

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

Penuel, W. R. & Means, B. (2004). Implementation variation and fidelity in an inquiry science program: Analysis of GLOBE data reporting patterns. *Journal of Research in Science Teaching, 41*, 294-315.

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

Princiotta, D., Flanagan, K. D., and Germino Hausken, E. (2006). Fifth Grade: Findings From The Fifth-Grade Follow-up of the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K). (NCES 2006-038) U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Putnam, R. T., Lampert, M., & Peterson, P. L. (1990). Alternative perspectives on knowing mathematics in elementary schools. *Review of Research in Education, 16*, 57-150.

Rittle-Johnson, B. & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology, 99*, 561-574.

Schulte, A. C., Easton, J. E., & Parker, J. (2009). Advances in treatment integrity research: Multidisciplinary perspectives on the conceptualization, measurement, and enhancement of treatment integrity. *School Psychology Review, 38*, 460-475.

Sfard, A. (2003). Balancing the unbalanceable: The NCTM Standards in the light of theories of learning mathematics. In J. Kilpatrick, G. Martin, & D. Schifter (Eds.), *A research companion for NCTM Standards* (pp. 353-392). Reston, VA: National Council of Teachers of Mathematics.

Smith, T., Cobb, P., Farran, D. C., Cordray, D. S., Munter, C., Green, S. E., Garrison, A., & Dunn, A. C. (2010, March). *Evaluating Math Recovery: Implications for policy and practice.* Paper presented at the annual meeting of the Society for Research on Educational Effectiveness, Washington, D. C.

Songer, N. B. & Gotwals, A. W. (2005, April). Fidelity of implementation in three sequential curricular units. In S. Lynch (Chair) & C. L. O'Donnell, *"Fidelity of implementation" in implementation and scale-up research designs: Applications from four studies of innovative science curriculum materials and diverse populations.* Symposium conducted at the annual meeting of the American Educational Research Association, Montreal, Canada.

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

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

U. S. Department of Education. (2006). *Education research request for applications - 84.305.* Retrieved April 24, 2006 from Institute of Education Sciences Web site: http://ies.ed.gov/funding/

Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology, 61*, 620-630.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock Johnson III test of achievement*. Itasca, IL: Riverside Publishing.

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

Wright, R. J., Martland, J., & Stafford, A. K. (2006). *Early numeracy: Assessment for teaching and intervention* (2nd ed.). London: Paul Chapman Publishing.

Wright, R. J., Martland, J., Stafford, A. K., & Stanger, G. (2006). *Teaching Number: Advancing children's skills and strategies* (2nd ed.). London: Paul Chapman Publishing.