Abstract Title Page

Not included in page count.

Title: Evaluating Math Recovery: Assessing the Causal Impact of Math Recovery on Student Achievement

Author(s): Thomas Smith, Paul Cobb, Dale Farran, David Cordray, Charles Munter and Alfred Dunn

Abstract Body

Limit 5 pages single spaced.

Background/context:

Description of prior research, its intellectual context and its policy context.

Children enter school at a wide range of mathematical abilities (Baroody, 1987; Dowker, 1995; Gray, 1997; Griffin & Case, 1999; Housasart, 2001; Wright, 1991, 1994a; Young-Loverage, 1989). A study conducted by Aunola, Leskinen, and Lerkkanen (2004) found that, in the absence of intervention, the initial gap in mathematics achievement continues to widen. Zill and West (2000) examined data from the U.S. Department of Education's Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), to describe the nature of this pre-K gap in mathematics achievement. On the one hand, they found that 20% of all kindergarten students in the U.S. in 1998 (approximately 780,000 students) were already beyond counting and reading single-digit numerals, and 4% (156,000) were even doing arithmetic. On the other hand, 42% (1.6 million) could not count up to 20 objects, and 6% (234,000) were unable to count even 10 objects.

Initial gaps in student achievement are persistent and in many cases widen as student progress in school (Aunola et al., 2004; Cockcroft, 1982). Duncan, Claessens, and Engel's (2004) analysis of ECLS-K indicated that pre-K mathematical ability is highly predictive of achievement at the end of first grade. Although teacher- and parent-reported social and emotional behaviors had standardized coefficients between .01 and .05, early mathematics abilities had coefficients in the .35 to .50 range for the subsequent first-grade data. Princiotta, Flanagan, and Germino Hausken's (2006) analysis of ECLS-K data revealed that achievement gaps are still prevalent in fifth grade. They found that 67% of students who scored in the top third in their kindergarten year did so again six years later; and that those among the lowest third in 1998 generally scored low in 2004.

Children's differing levels of mathematical ability when they enter school are related to multiple factors. Children who are less ready for school typically come from families of low SES status, are of racial or ethnic minority backgrounds, have parents who do not speak English in the home, or possess disabilities (Alexander & Entwistle, 1998; Barton, 2003; Berends, Lucas, Sullivan, & Briggs, 2005; Cahalan et al., 2006; Chen, 2005; Crosnoe, 2005; Dahlstrom, 2005; Fuson, Smith, & Lo Cicero, 1997; Griffin, Case, & Siegler, 1994; Griffin, 2004; Marchand, Pickreign, & Howard, 2005; Vandivere, 2004; Walker, 2006; Wilms, 1986). However, it is important to note that ethnicity per se does not predict success in mathematics once it is adjusted for other factors (Thomas, 2000). One potential explanation for the persistence of the initial achievement gap focuses on inequities in the educational opportunities across ethnic, racial, and socioeconomic groups. In this regard, Oakes (1990) found that "[s]chools with large concentrations of low-income and minority students [tend to] offer fewer classroom conditions that are likely to promote active engagement in mathematics and science learning—such as opportunities for hands-on activities and time working with the teacher" (p. 101).

The research findings we have reviewed thus far indicate that differences in early mathematical abilities are relatively stable and can lead to differentiated instruction in the later years of elementary school and in middle school. The findings emphasize the pressing need to equip schools with effective methods for closing the pre-K gap (McWayne, Fantuzzo, & McDermott, 2004).

Purpose / objective / research question / focus of study:

Description of what the research focused on and why.

Our goal was to evaluate the potential of Math Recovery (MR), a pullout, one-to-one tutoring program that has been designed to increase mathematics achievement among low-performing first graders, thereby closing the school-entry achievement gap and enabling participants to achieve at the level of their higher-performing peers in the regular mathematics classroom.

Specifically, our research questions were as follows:

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?

Setting:

Description of where the research took place.

The two-year evaluation of Math Recovery 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.

Population / Participants / Subjects:

Description of participants in the study: who (or what) how many, key features (or characteristics).

Students were selected for participation at the start of first grade based on their performance on MR's screening interview and follow-up assessment interview. The screening is designed to select the lowest achieving first graders (25th percentile and below) in terms of math achievement. The number of students eligible for tutoring ranged from 17 to 36 across the 20 schools. The number of study participants before attrition totaled 517 in Year 1 and 510 in Year 2, of which 172 received tutoring in Year 1 and 171 received tutoring in Year 2. Approximately 50% of participants were males, 48% were non-white and 48% received free or reduced lunch.

We recruited 18 teachers to receive training and participate as MR tutors from the participating districts—all of whom had at least two years of classroom teaching experience. Sixteen of the tutors received half-time teaching releases to serve one school each; two of the tutors received full-time teaching releases to serve two schools each. All tutoring positions were underwritten by their respective school districts.

Intervention / Program / Practice:

Description of the intervention, program or practice, including details of administration and duration.

MR consists of three components: 1) tutor training, 2) student identification and assessment, and 3) one-to-one tutoring. The first component of the MR program, tutor training, involves 60 hours of instruction provided by an MR leader. The goal of this training is to support tutors' in learning new practices for clinical assessment and intervention teaching in which they use the Learning Framework and the Instructional Framework to adjust instruction based on cognitive evaluations of student responses.

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. 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.

Research Design:

Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

The structure of the MR program allowed us to use the fact that two thirds of the participating students will have their treatment 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 has not yet begun and a small number of students who are on a "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, we can establish the essential characteristics of an experimental design: a comparison of students' change in mathematics achievement during their 12 weeks of participation in MR to the gains they would have made if they not participated in the intervention.

In each year (2007-08 and 2008-09 academic years), three eligible students from each school were randomly assigned to a tutoring cohort with a different start date (i.e., Cohort A—September, Cohort B—December, Cohort C—March) or to the "waiting list" for MR. In both years students on the randomly ordered waiting list were selected to join an MR tutoring cohort if an assigned participant left their school or were deemed "ineligible" due to a special education placement.

Data Collection and Analysis:

Description of the methods for collecting and analyzing data.

Each of the students participating in the study were assessed using alternating forms of the Applied Problems, Quantitative Concepts, and Fluency subtests of the Woodcock Johnson III Achievement tests (WJ III) subtests, as well as the MR proximal instrument designed in consultation with the program developers, at the start of the study and when each cohort entered or exited tutoring in December, March, and May. 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.

Our research design allowed us to describe and compare the growth trajectories of treatment and control cohorts across the whole school year, punctuated at the end of each 11-week period by the students completing MR tutoring. We used the estimated growth rate of Cohorts 1B and 1C prior to receiving treatment, as well as the estimated growth rates of students on the wait list who did not receive the MR intervention, to estimate a counterfactual to the growth rate of MR participants. At the end of each of these intervals, a given study participant has one of three statuses: not yet received any MR, just completed MR tutoring, or is post MR. We made different comparisons within this scheme to determine mathematics achievement outcomes immediately at the end of an MR session relative to students who had not received the MR intervention (to test the treatment effect) and outcomes 12 or 24 weeks after completing MR to those who have just completed MR (to test whether MR gains are maintained after the end of treatment).

To estimate these growth trajectories, we used a 3-level hierarchical linear growth models (Raudenbush and Bryk, 2002; Singer and Willett, 2002) with repeated observations of WJ III scores or MR proximal scores indexed by time, time since starting MR, and time since completing MR at level 1, student level demographics at level 2 (e.g., gender, minority status), and school characteristics at level 3. To assess whether gains made in MR tutoring are maintained after the tutoring is completed, a time varying covariate (POSTMRTIME$_{ijt}$) that counts the number of days after a student completes MR. The level 1 equation looks like:

$$WJIII_{ijt}=\pi_{0j}+ \pi_{1j}(Time)_{ijt}+ \pi_{2ij}(MRTIME)_{ijt} + \pi_{3ij}(POSTMRTIME)_{ijt} + \varepsilon_{ijt}$$

Thus, the coefficient π_{2ij} on MRTIME$_{ijt}$ can be interpreted as the treatment effect—the additional daily learning associated with participation in MR relative to non-participants and cohorts who have not yet begun the tutoring program. The coefficient π_{3ij} on POSTMRTIME$_{ijt}$ can be interpreted as the additional daily learning for participants after completing MR compared to their rate of learning when participating in MR tutoring. Although the results presented here are only for the first year cohort in this study, the paper presented at SREE will include end of second grade data for Cohort 1 and end of first grade data for cohort 2. We are particularly interested in testing the hypothesis that the gains made from participation in MR are maintained through the end of second grade.

Findings / Results:
Description of main findings with specific details.

The first year results show a small to moderate effect of participation in MR on WJ III scores and moderate to large effects on the MR proximal assessments. Specifically, differences in the end of first grade mean scores on the WJ III subtests between students selected for tutoring and those on the waitlist ranged in effect size from .21 on the quantitative concepts scale to .28 on the applied problems scale (all differences statistically significant at the $p<.05$ level). Effect sizes on the MR 1.1 screening assessment ranged from .34 on the forward number sequence scale to .92 on the arithmetic strategies measure. These results compare favorably to those reviewed recently by Slavin and Lake (2006), including several cooperative learning programs that had median effect sizes of at least +0.30 in studies using randomized experimental or randomized quasi-experimental designs, including Class wide Peer Tutoring (.33), Student Team Learning (.19-.60), and TAI Math (.28-.38). A meta-analysis of 52 studies on the relationship between tutoring and student achievement (Cohen, Kulik, and Kulik, 1982), however, found average effect sizes greater than .40—higher than MR effects on the WJ III measures but lower than effects on some of the more proximal assessments.

Results from the growth models show increases in mathematics achievement for MR participants across all assessments during the tutoring period (with p<.05 in each case), although this growth rate tends not to be maintained after completion of MR. For example, on the applied problems subscale of the WJ III, MR participants gained .063 points per day, on average, during tutoring while students on the wait list for MR gained .038 points per day across the same time period. After exiting MR tutoring, however, participants' growth trajectories reverted back to pre-tutoring rates—a rate of .033 per day on the applied problems subscale. The pattern of strong gains pre to post tutoring, with regression towards the growth trajectories of non-participants was consistent across assessments.

By November 2009, we will have completed processing and cleaning of the Year 2 data for this study, allowing us to test whether MR tutors are more effective in their second year of tutoring than their first. We will also test differences in both the WJ III scores and a second grade version of the MR proximal assessment to determine whether the gains made by participants in first grade are maintained through the end of second grade.

Conclusions:
Description of conclusions and recommendations based on findings and overall study.

The findings of this study have theoretical, practical, and policy significance. Practically, the positive causal effect of MR tutoring demonstrates that programs that are diagnostic rather than scripted in nature can overcome fidelity concerns and have an impact on student early mathematics performance. Theoretically, our findings indicate that investing in tutors' knowledge of student reasoning and pedagogical content knowledge can pay off in terms of improvement in student's mathematical learning, particularly if tutors use carefully designed tools such as the MR Learning and Instructional Frameworks that codify and schematize this knowledge. With regard to policy, our finding that the MR program can reduce some of the pre-K mathematics achievement gap provides an initial indication that the cost of the program per student might be justified, although further work is needed to understand why initial gains made by participants appear to diminish after tutoring ends. It is possible that the forms of arithmetic reasoning that MR develops needs to be further supported in the regular classroom to see the full benefit of this form of tutoring. Longitudinal studies that track MR students and their initially higher performing peers until the end of elementary school are needed to address this question adequately.

.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

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

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

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

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. & 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.

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

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

Dowker, A. (1995). Children with specific calculation difficulties. *Links 2*(2), 7-11.

Duncan, G. J., Claessens, A., & Engel, M. (2004). *The contribution of hard skills and socio-emotional behavior to school readiness*. Retrieved October 26, 2006, from http://www.northwestern.edu/ipr/people/duncanpapers.html.

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., Smith, S. T., & Lo Cicero, A. M. (1997). Supporting Latino first graders' ten-structured thinking in urban classrooms. *Journal for Research in Mathematics Education, 28*(6), 738-766.

Gray, E. M. (1997). Compressing the counting process: Developing a flexible interpretation of symbols. In I. Thompson (Ed.), *Teaching and learning early numbers* (pp. 63-72). Buckingham: Open University Press.

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

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

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

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

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

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.

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

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

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

Young-Loveridge, J. M. (1989). The development of children's number concepts: The first year of school. *New Zealand Journal of Educational Studies, 24*(1), 47-64.

Appendix B. Tables and Figures

Not included in page count.

Abstract Title Page

Not included in page count.

Title: Evaluating Math Recovery: Investigating Tutor Learning

Author(s): Sarah Elizabeth Green and Thomas Smith

Abstract Body

Limit 5 pages single spaced.

Background/context:

Description of prior research, its intellectual context and its policy context.

The success of Math Recovery, a constructivist, adaptive one-on-one tutoring intervention depends heavily on the skill and knowledge of the teachers who are selected to deliver it. MR is adaptive in that a key component of the program is that the tutor is expected to adjust instruction to the current level of a student's thinking at any given point in time. This makes the practice of conducting the tutoring far more demanding than many more scripted or prescribed interventions.

In the course of Math Recovery training, tutors are asked to think about mathematics and student learning in ways that are not typical in many US classrooms. Therefore, they are being asked to learn a new of way of teaching mathematics to students. The complex learning demands on new tutors are similar to the learning demands on classroom teachers attempting to implement ambitious reform mathematics instruction. Teachers learn from various sources in addition to their formal professional development, such as learning in and from practice (Franke & Kazemi, 2001). The literature has shown that teachers' learning in practice often results from gaining access to students' thinking and reasoning and having to reorganize their own understanding as a result (Franke, Carpenter, Fennema, Ansell, & Behrend, 1998). Since a key component of MR is uncovering, diagnosing, and building upon students' thinking the potential for this type of generative learning in practice is substantial. Therefore, while it may seem that selecting tutors at the outset of the adoption of MR for their knowledge and skills would be important, the extent to which tutors can learn from enacting MR and gain the necessary knowledge as a result is still unknown. This analysis begins a line of investigation that attempts to understand the relationships between teachers' knowledge and teachers' learning in practice.

One potential area where tutors may grow in their knowledge as a result of MR is their mathematical knowledge for teaching (MKT). While MKT may generally be important for good mathematics teachers (Ball, Lubienski, & Mewborn, 2001; Ball, Thames, & Phelps, 2008) it is particularly central to MR. MKT is knowing math in a specialized way that is particular to the profession of teaching. For instance, while average people competent in mathematics need to know how to add multi-digit numbers, teachers also need to know why the conventions for addition work, what are typical ways that students approach these kinds of problems, common errors they make, non-conventional methods that will work, which methods will be best built on later in their mathematical learning, etc. In MR, tutors are consistently assessing students' current methods for solving problems and determining how to build on their current understanding, which implicates their MKT as an important aspect of the knowledge needed for good MR tutors. Researchers at the University of Michigan have developed an assessment of this type of math knowledge, the Learning Mathematics for Teaching (LMT) assessment (Hill, Schilling, & Ball, 2004). In this analysis we use tutors' performance on the LMT as a way to understand how this kind of knowledge changes as a result of being a MR tutor.

Another important area of knowledge for MR tutors is their knowledge of the MR Learning and Instructional Frameworks in Number (LFIN and IFIN, respectively). MR training is dedicated in large part to tutors learning to understand and use these frameworks as a part of their practice as MR tutors. The frameworks lay out developmental trajectories for students in early number learning and suggest instructional activities to support students at various points along those trajectories. A tutor's ability to understand and use the frameworks is key to their implementation of the tutoring program. Again, it may be that as tutors gain experience with different students and their thinking their understanding and skill in using the frameworks becomes more complete. Anticipating and understanding students' thinking is central to both MKT and understanding the MR frameworks. This may mean that tutors who have more MKT can learn to understand and enact MR tutoring more quickly and effectively. This analysis uses tutors scores on a test of their knowledge of the frameworks (TKA) to investigate this aspect of changing knowledge as a result of gaining experience as a MR tutor and the relationship to MKT.

There are two typical ways to enhance the skill and knowledge of teachers: hiring and professional development. While selecting teachers based on skills and knowledge seems logical, this may not always be practical for districts and schools hoping to adopt a new program. Absent direction from researchers or program developers districts may not know what qualities and skills are necessary and if they do know they may not be able to locate that type of expertise in their local context. Additionally, some of the knowledge and skill needed to deliver an intervention such as Math Recovery is specialized to the intervention itself. In this paper we investigate these possibilities, as well as a third, that tutors learn from the practice of tutoring itself. The findings from this analysis have potential implications for policy and implementation of MR as well as for future studies of this and like interventions.

Purpose / objective / research question / focus of study:
Description of what the research focused on and why.

The goal of the overall study was to evaluate the potential of Math Recovery (MR), a pullout, one-to-one tutoring program, to increase mathematics achievement among low-performing first graders, thereby closing the school-entry achievement gap by enabling them to achieve at the level of their higher-performing peers in the regular mathematics classroom.

An additional purpose of the study is to inform the design of future effectiveness and scale-up studies, should they be warranted, as well as policy decisions regarding adopting the program and selecting tutors. In order to achieve these purposes, we used two measures of tutor knowledge, the LMT and a developer created test covering the understanding and use of the MR frameworks.

This analysis answers the following research questions regarding tutor knowledge:

1. Were there initial differences in the MKT of the tutors by site?
2. Were there differences in the uptake of MR training by site? Did tutors' knowledge of the frameworks differ by site at the start of the experiment?
3. Do initial differences in tutor knowledge (both MKT and knowledge of MR frameworks) persist as tutors gain experience with MR and learn through practice?

4. How do differences in tutor knowledge and tutor learning relate to student outcomes?[*]

Setting:

Description of where the research took place.

The two-year evaluation of Math Recovery 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. The newness of MR to the 18 tutors (two tutors worked at more than one school) makes investigating the change in knowledge over time particularly relevant for suggesting potential policy implications in new adoptions of the program. Each state was a different "site" in the context of this analysis in that they received different initial training experiences and had different ongoing support for their learning of the program.

.

Population / Participants / Subjects:

Description of participants in the study: who (or what) how many, key features (or characteristics).

Students were selected for participation at the start of first grade based on their performance on MR's screening interview and follow-up assessment interview. The screening is designed to select the lowest achieving first graders (25th percentile and below) in terms of math achievement. The number of students eligible for tutoring ranged from 17 to 36 across the 20 schools. The number of study participants before attrition totaled 517 in Year 1 and 510 in Year 2, of which 172 received tutoring in Year 1 and 171 received tutoring in Year 2.

The participating districts hired 18 teachers to receive training and participate as MR tutors—all of whom had at least two years of 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 and thus were full time tutors.

Intervention / Program / Practice:

Description of the intervention, program or practice, including details of administration and duration.

MR consists of three components: 1) student identification and assessment, 2) one-to-one tutoring, and 3) tutor training. In the first 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 second 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. The tutor's selection of tasks for sessions with a particular child is initially informed by the

[*] This question will be addressed in the full paper, though it is not included in the remainder of this proposal.

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 third component of the MR program, tutor training, involves 60 hours of instruction provided by an MR leader. The goal of this training is to support tutors' in learning new practices for clinical assessment and intervention teaching in which they use the LFIN and the IFIN to adjust instruction based on cognitive evaluations of student responses. Training was conducted at each site by MR expert trainers for five days during the summer and with five additional days of follow up throughout the first two months of tutoring. Additionally, the tutors had support for site coordinators who met with them on a monthly basis.

Research Design:

Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

The larger evaluation study was a randomized field trial. 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 the 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 were deemed "ineligible" due to a special education placement. The number of study participants totaled 517 in Year 1 and 510 in Year 2, of which 172 received tutoring in Year 1 and 171 received tutoring in Year 2.

To study teacher knowledge and learning, we assessed teachers at three time points with instruments discussed previously, the LMT and the TKA, to determine whether there were differences in the uptake of training across sites. Teachers at the two sites were trained somewhat differently as reported by the site coordinators and MR experts. Some of the differences in training were attributed to initial perceived differences in teachers' MKT. Therefore, while this study does not have a control group, the differences between the groups provide variation for us to attempt to understand some of the conditions of implementing tutoring that effect tutor learning. It is our hop to uncover important factors of tutor learning to study more rigorously in future studies.

Data Collection and Analysis:

Description of the methods for collecting and analyzing data.

The tutors were assessed using the LMT assessment to measure their MKT and the TKA to asses their knowledge of the MR frameworks. The assessments were given at three time points: end of MR training, end of year 1, and end of year 2. These assessments were scored and double-entered into a central database for further analysis.

This analysis uses one-way analysis of variance (ANOVA) where the predictor is training site membership. There were two main training sites. Therefore, in order to answer the questions that are a focus of this investigation, ANOVA is used to test for a difference between means of the groups on both the TKA and the LMT at time 1 and at time 3.

In the full paper, we include hierarchical linear models (Raudenbush and Bryk, 2002) to model growth of tutor knowledge and also use this analysis method to link student outcomes (see proposal 1 for a full description) to learning rates of tutors during tutoring.

Findings / Results:
Description of main findings with specific details.

Initially, there is a significant difference between groups at time 1 on the LMT (F=7.81, p = 0.013) and on the TKA (F=15.18, p=0.0013). This indicates that the tutors in site A were initially more knowledgeable in their MKT and also learned more about the MR frameworks from the initial MR training. This confirms the reports of the site coordinators and MR experts who conducted the training. However, at the end of the study there are no longer significant differences between these groups on either measure (LMT: F=3.55, p=0.08 & TKA: F= 1.36, p=0.26). Plotting the means of the groups on both measures over the three time points shows that both groups increased their mean scores on these measures over the three time points and therefore the lack of difference between groups at the end of the study is not due to tutors at site A decreasing in knowledge, but rather a steeper increase in knowledge at site B. In the full paper growth models are used to discuss this pattern of knowledge growth further.

Conclusions:
Description of conclusions and recommendations based on findings and overall study.

These results have two implications for policy and for future studies of this intervention. First, tutors who had higher MKT at the outset also had higher scores on the TKA (r = 0.5, p= 0.03). Since, these tutors are all new to MR, this suggests tutors with more math knowledge for teaching may learn more from the initial MR training, potentially making them better choices for tutoring early on. Second, the initial differences did not persist between groups after two years of tutoring experience which suggests that tutors can and do grow in their understanding of the MR frameworks and also in their math knowledge for teaching through their MR tutoring practice. This is likely related to repeated attempts to understand students' thinking. Students' thinking and solution method is a key aspect of MKT and also an important part of using the MR frameworks with understanding. Understanding the exact mechanism for how tutors learn from the practice of tutoring students is an issue for research. An implication for policy and adoption of MR is that while initially tutors may struggle in their knowledge of MR, time and experience with the program will likely increase their knowledge of the program over time.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Ball, D. L., Lubienski, S. T., & Mewborn, D. S. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. *Handbook of research on teaching, 4*, 433-456.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389-407.

Franke, M. L., Carpenter, T., Fennema, E., Ansell, E., & Behrend, J. (1998). Understanding teachers' self-sustaining, generative change in the context of professional development. *Teaching and Teacher Education, 14*(1), 67-80.

Franke, M., & Kazemi, E. (2001). Teaching as learning within a community of practice: Characterizing generative growth. In T. Wood, B. C. Nelson & J. Warfield (Eds.), *Beyond classical pedagogy in elementary matheamtics: The facilitative nature of teaching* (pp. 47-74). Mahwah, NJ: Erlbaum.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *The Elementary School Journal, 105*(1), 11-30.

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

Appendix B. Tables and Figures

Not included in page count.

Abstract Title Page

Not included in page count.

Title: Evaluating Math Recovery: Measuring Fidelity of Implementation

Author(s): Charles Munter, Anne Garrison, Paul Cobb, and David Cordray, Vanderbilt University

Abstract Body

Limit 5 pages single spaced.

Background/context:

Description of prior research, its intellectual context and its policy context.

One of the primary 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 that 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. Claims of treatment effectiveness may be unjustified and invalid unless the degree to which programs are implemented as intended is defined and assessed. However, despite this emphasis on measuring implementation fidelity, recent reviews of studies in school settings have illustrated that many inconsistencies and omissions in measuring fidelity exist (Dusenbury, 2003; O'Donnell, 2008). Furthermore, little is known regarding the feasibility of conducting studies of implementation fidelity of unscripted interventions, where measuring fidelity first requires the identification and operationalization of complex, subtle facets of the intervention (Cordray & Pion, 2006).

Purpose / objective / research question / focus of study:

Description of what the research focused on and why.

In this paper, we describe a case of measuring implementation fidelity within an evaluation study of Math Recovery (MR), a pullout tutoring program aimed at increasing the mathematics achievement of low-performing first graders, thereby closing the school-entry achievement gap by enabling them to achieve at the level of their higher-performing peers in the regular mathematics classroom. Two research questions guided the conduct and analysis of the larger 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 analysis reported in this paper follows from a third question: 3) To what extent does fidelity of implementation influence the effectiveness of MR?

Math Recovery one-to-one tutoring is diagnostic in nature and focuses instruction at the current limits of each child's arithmetical reasoning. The tutor's selection of tasks for sessions with a particular child is initially informed by an assessment interview and then by ongoing assessments based on the student's responses to prior instructional tasks. Therefore, measuring fidelity in this case is not as simple as monitoring adherence to a script, but requires assessing the extent to which a tutor's instruction is consistent with the complex practice of attuning instruction to a child's current level of mathematical reasoning.

Our goals were to both measure the extent to which the program was implemented as intended, and, eventually, to link the measures to student outcomes. Determining the extent to which the tutoring is enacted as intended requires an explication of 'good' tutoring as defined by the developers and systematically evaluating tutors' practices against that ideal. However, we also go beyond MR's notion of 'good' tutoring by looking for instances of "positive infidelity" (Cordray, 2009) within tutoring sessions, including aspects of instruction identified in mathematics education research literature as being effective in supporting students in learning mathematics with understanding, but not included in the MR model. Thus, we view studies of implementation fidelity as potential sources for refining theory and program design.

Setting:
Description of where the research took place.

The two-year evaluation of Math Recovery 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.

Population / Participants / Subjects:
Description of participants in the study: who (or what) how many, key features (or characteristics).

Students were selected for participation at the start of first grade based on their performance on MR’s screening interview and follow-up assessment interview. Eighteen teachers were recruited to receive training and participate as MR tutors from the participating districts—all of whom had at least two years of 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. All tutoring positions were underwritten by their respective school districts.

Intervention / Program / Practice:
Description of the intervention, program or practice, including details of administration and duration.

Math Recovery 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.

Guiding the fidelity assessment were what we, in collaboration with program developers, determined to be the unique aspects of Math Recovery tutoring as compared to typical tutoring: (a) the tutor’s ongoing assessment of the child’s thinking and strategies (both reflective assessment between tutoring sessions and in-the-moment assessment); and (b) the tutor’s efforts to provide instruction within the child’s zone of proximal development.

Research Design:

Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

The larger evaluation study was a randomized field trial. 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 were deemed "ineligible" due to a special education placement. The number of study participants totaled 517 in Year 1 and 510 in Year 2, of which 171 received tutoring in Year 1 and 172 received tutoring in Year 2.

In this paper, however, we report primarily on the process of determining the reliability and validity of fidelity indices. At the outset, we consulted program developers to identify key implementation components (Fixsen *et al*., 2005) and initial schemes for measuring those constructs. Occurring across 20 hours during three days, this consultation was no trivial task. Although the program developers had a relatively well-articulated theory, no measures had previously been established, and doing so required meeting the challenge of bringing developers' and researchers' perspectives to a consensus. The research team finalized the instruments through an iterative refinement process, based on multiple rounds of video coding and grounded in MR's guiding principles, eventually establishing adequate (90%) agreement.

Five coders, each with experience in either elementary classroom instruction or video coding (or both), were hired and received two kinds of training, including a five-day session led by MR experts 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 (led by members of the evaluation team). 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 do administer the MR assessment with first-grade students; (d) an introduction to the materials typically utilized in MR instruction; and (e) direction on coordinating the Learning Framework with the Instructional Framework. The rationale for providing such extensive training on the program itself was that coders' work would be more likely to faithfully represent the spirit of the program if they had firsthand experience in examining its underlying theory and in employing its fundamental tools.

The initial four-day coding training included (a) an introduction to the research team's operationalizations of the core implementation components of MR, including the instruments the research team had developed; (b) multiple rounds of collective video coding, during which we paused to discuss 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 agreement reached an adequate level (80%). Throughout this final, four-week phase, we met weekly with the coding team to further refine, define and operationalize the aspects of MR that they were attempting to code. Thus, early on, coders' feedback was important in increasing the feasibility of MR fidelity assessment.

As stated above, consistent with typical MR practice, all assessment and tutoring sessions were video-recorded. Approximately 20% of the tutoring cycles were randomly selected to be assessed for fidelity of implementation—one student per cycle per tutor (a total of 108 students across all 18 tutors and all 6 cycles). For each student selected, coders assessed the fidelity with which the initial assessment and 12 instructional lessons were conducted. To select the lessons for coding, we divided the total number of lessons received into six equal blocks and randomly selected two lessons from each block. This totaled 216 assessments and 1,296 tutoring sessions coded for 108 students.

For purposes of external validation, a subset of assessment and tutoring sessions were sent to 30 MR experts, who rated the tutoring practices based on their own notions of high-quality MR practice. Eight assessment sessions and twelve instructional lessons were selected to represent range of scores on indices of implementation fidelity as determined by our coding schemes. The MR experts were asked to determine the extent to which tutors enacted MR as intended, using their own criteria. Specifically, for both assessments and instructional lessons, they were asked to 1) rank, from highest to lowest, the tutors' enactments of MR as intended, and 2) indicate which of four categories they would place each video: *excellent*, *good*, *fair* or *poor*. Each video was labeled with a pseudonym for reference, and the MR experts remained blind to the research team's instruments and assessment criteria until after the validity study was complete.

Data Collection and Analysis:

Description of the methods for collecting and analyzing data.

Guided by the unique aspects of Math Recovery tutoring listed above (i.e., the tutor's ongoing assessment of the child's thinking and efforts to provide instruction within the child's zone of proximal development), our goal in assessing implementation fidelity was to answer a set of key questions regarding tutors' assessment and instruction: (a) Was the initial assessment done? If so, was it done correctly? (b) In instructional lessons, did the tutor choose procedures (i.e., sets of related tasks) that were in the child's zone of proximal development (according to the MR Frameworks)? (c) Did the tutor utilize/implement the procedures/tasks well?

Regarding the first question, we identified two possibilities for breakdown: the tutor might have 1) presented the incorrect assessment tasks (or tasks that were misaligned with those printed in the assessment), or 2) used poor judgment in interpreting the results (i.e., assigned a profile to the student that conflicted with our external assessment of the child's current understanding). For each of these we defined what constituted a *minor error*, a *major error*, or *no error*.

To answer the second question, regarding tutors' *choice* of procedures, coders first viewed up to three previous tutoring sessions to locate the child's thinking at that point on the MR Learning Framework, and then determined whether the tutor's choice of procedures matched 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? Often tutors utilized procedures as described in the MR handbook, but when they incorporated procedures from other sources, coders located those procedures on the Instructional Framework based on the procedure's focus (e.g., arithmetical strategies, number word sequences, etc.), and the level of difficulty of the tasks within the procedure, including number range and the extent to which the tasks were scaffolded.

Lastly, to answer the question pertaining to tutors' *implementation* of tasks (within procedures), coders examined the extent to which tutors followed established "rules" within the MR program (e.g., things a tutor is supposed to do, or prohibitions). For example, tutors are expected to consistently solicit students' strategies for solving problems (if the strategy is not already visible), and are expected to avoid merely eliciting particular behaviors.

After four weeks of refinement work (described above), 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?) Therefore, the evaluation team 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 MR’s unique aspects listed above. Therefore, video coding focused on instances of activities aimed at supporting students in developing more sophisticated *strategies*, rendering the fidelity assessment process more tractable without sacrificing any attention to core implementation components.

Findings / Results:
Description of main findings with specific details.

Throughout the coding process (after the initial refinement phase), coders maintained an average percent agreement of 0.80. Furthermore, MR experts’ ratings validated our coding schemes, with sufficiently high correlations between their ratings and those based on fidelity indices.

Conclusions:
Description of conclusions and recommendations based on findings and overall study.

Our findings suggest it is possible to create a reliable instrument to measure implementation fidelity for differentiated interventions—an endeavor that has, heretofore, been largely avoided in evaluations of educational interventions. Many potentially high-quality interventions are un-scripted, instead relying on teacher knowledge and professional development, requiring considerable differentiation by implementers. As we work to rigorously evaluate such programs, we need to develop reliable fidelity measures that are both feasible and true to program components, so that evaluators can adequately link measures of treatment integrity to outcomes, to more accurately determine the relative strength of interventions (Cordray & Pion, 2006). This paper outlines the development and use of one such measure as a case of how such fidelity instruments might be developed and used in the future. Critical aspects of the process included 1) the identification of the core implementation components of the intervention (Fixsen *et al*., 2005); 2) close work with program developers to operationalize those components; 3) training of coders in both the program itself and the coding schemes/process; and 4) collaborating with the coding team to further refine operationalizations and coding decisions, to strike a balance of feasibility and adherence to program components.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

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.

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.

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., Blasé, 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).

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.

Appendix B. Tables and Figures

Not included in page count.