Abstract Title Page

Title:

Evaluating Math Recovery: Implications for Policy and Practice

Author(s):

Thomas Smith (with Paul Cobb, Dale Farran, David Cordray, Charles Munter, Sarah Green, Annie Garrison, and Alfred Dunn), Vanderbilt University

Abstract Body

Background/context:

This presentation focuses on an initial evaluation study 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. Following Cordray and Morphy (2009), our goal was not merely to assess whether MR works, for whom, under what circumstances. We also attempted to understand how and why the program works to produce particular outcomes (cf. Clements, 2007). In addition, we illustrate that assessments of implementation fidelity can help identify aspects of an intervention that need improving. Assessments of implementation fidelity in turn require that 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" (Hulleman & Cordray, 2009, p. 90).

The rationale for the evaluation study is grounded in the well-documented finding that 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. Furthermore, 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, and Princiotta, Flanagan, and Germino Hausken's (2006) analysis of the same data set 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. Taken together, these findings emphasize that identifying effective methods for closing the pre-K gap is an pressing policy concern (McWayne, Fantuzzo, & McDermott, 2004). Using an experimental design, we assessed the effectiveness of MR in improving mathematics achievement of low performing first grade students and examined whether gains made in first grade were maintained through the end of second grade.

As we clarify below, MR tutoring is a demanding form of practice in which tutors are expected to adjust instruction to the current level of a student's thinking at any given point in time. The teacher development literature suggests that teachers can learn in the context of their practice, often as they attempt to understand students' reasoning and adjust instruction accordingly (Franke, Carpenter, Fennema, Ansell, & Behrend, 1998). In other words, the effectiveness of MR tutoring may improve as the tutors gain additional experience with the intervention. Thus, although it might be important to select tutors based on their initial knowledge and skills, it is also essential to consider the extent to which tutors can develop the necessary knowledge as they enact MR. This necessary knowledge includes mathematical knowledge for teaching (MKT) (Hill, Rowan, & Ball, 2005). MKT denotes a form of mathematical knowledge that is specific to problems and decisions that arise in the practice of teaching. A priori, MKT appeared to be central to MR because tutors are expected to assess and build on students' current reasoning. A second aspect of the necessary knowledge concerns tutors' knowledge of the MR Learning and Instructional Frameworks in Number (LFIN and IFIN, respectively). A primary goal of MR training is to enable tutors to understand these frameworks, and to use them in their tutoring practice. 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 therefore integral to effective implementation of MR. Thus we examine both the impact of tutors initial MKT, LFIN, and IFN knowledge on their tutoring effectiveness and whether increases in tutors' knowledge in these areas over the course of the study is associated with increasing effectiveness in their tutoring.

It is widely acknowledged that claims of treatment effectiveness may be unjustified and invalid unless the degree to which programs are implemented as intended is defined and assessed (Dusenbury, 2003; O'Donnell, 2008). However, little is known regarding the feasibility of assessing the implementation fidelity of unscripted interventions such as MR, where measuring fidelity requires the identification and operationalization of complex, often implicit facets of the intervention (Cordray & Pion, 2006). As part of the study, we measured the implementation fidelity of MR tutoring and will eventually link the measures to student outcomes.

Purpose / objective / research question / focus of study:

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?
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. To what extent does fidelity of implementation influence the effectiveness of MR?

Setting:

The two-year evaluation of Math Recovery was conducted in 20 elementary schools (five urban, ten suburban, and five rural) from 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:

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.

We recruited 18 teachers to receive training and participate as MR tutors. 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.

Intervention / Program / Practice:

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 MR Learning Framework and the Instructional Framework to adjust instruction based on cognitive evaluations of student responses. The tutors participating in the study were trained by MR personnel at two sites in different states.

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 or 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:

The structure of the MR program allowed us to use the fact that two thirds of the participating students 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 could establish the essential characteristics of an experimental design.

To study teacher knowledge and learning, we assessed tutors at three time points with a measure of their MKT developed by Ball, Hill and colleagues (e.g., Hill, Ball, & Schilling, 2008), and a measure of their knowledge of the MR Learning and Instructional Frameworks constructed in collaboration with the developers of the MR program.

As part of their standard practice, MR tutors video-record all tutoring sessions in order to plan for subsequent sessions. In the presentation, we will describe both the iterative process by which we developed an instrument for assessing the fidelity of implementation of MR tutoring, and the process by which we validated the instrument by comparing ratings on a subset of the video-recorded tutoring sessions with the assessments of 30 MR experts. We will also describe how we trained coders until agreement reached an adequate level (80%). The coders are currently coding a randomly selected 20% of the video-recorded tutoring sessions.

Data Collection and Analysis:

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 (WJ III) subtests, as well as the MR proximal instrument, an assessment based on the learning framework that we 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. To estimate these growth trajectories, we used 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 that counts the number of days after a student completes MR. 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.

The tutors were assessed using the externally validated LMT assessment to measure their MKT and an internally designed Tutor Knowledge Assessment (TKA) to assess their knowledge of the MR frameworks. The assessments were given at end of MR training, end of year 1, and end of year 2. The analysis of these data used a one-way analysis of variance (ANOVA) where the predictor was training site (as noted above, there were two training sites). ANOVA was used to test for a difference between means of the two groups on both the TKA and the LMT at time 1 and at time 3.

Findings / Results:

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. 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 proximal measures 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). However, a meta-analysis of 52 studies on the relationship between tutoring and student achievement (Cohen, Kulik, and Kulik, 1982) 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.

With regard to the tutors, there is a significant difference between the two training sites at time 1 on the LMT ($F=7.81$, $p = 0.013$) and on the TKA ($F=15.18$, $p=0.0013$), indicating that the tutors at one site were initially more knowledgeable in their MKT and also learned more about the MR frameworks from the initial MR training. However, at the end of the study there were no significant differences between these groups on either measure (LMT: $F=3.55$, $p=0.08$ & TKA: $F= 1.36$, $p=0.26$). This lack of difference between groups at the end of the study was due to a steeper increase in knowledge at the second site. We will present data on the relationship between tutor LMT and TKA scores during the presentation.

Conclusions:

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. Our findings therefore 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. 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.

The results concerning the tutors’ mathematical knowledge and knowledge of the MR Frameworks 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, suggesting that 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, indicating 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. As a consequence, any limitations in the MR program indicated by the second-year findings cannot be attributed to tutors’ lack of knowledge of the Frameworks. An implication for policy and adoption of MR is that while initially tutors might struggle to learn the MR Frameworks, their knowledge of the both the Frameworks and their MKT will likely improve with experience with the program.

The findings from our fidelity study suggest that it is possible to create a reliable instrument to measure implementation fidelity for differentiated interventions, an endeavor that has typically been largely avoided in evaluations of educational interventions. Many potentially high-quality interventions are un-scripted and instead rely on teacher knowledge and professional development. Because the fidelity measures are reliable and true to program components, we will be able to link measures of treatment integrity to outcomes, further clarifying how and why MR to produce particular outcomes (Cordray & Pion, 2006). Critical aspects of the process included 1) the identification of the core implementation components of the intervention (Fixsen *et al*., 2005); 2) operationalization of those components; 3) training of coders in both the program itself and the coding schemes; and 4) collaborating with the coding team to further refine coding decisions.

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.

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

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.

Cordray, D. S., & Morphy, P. (2009). Research synthesis and public policy. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (pp. 473-493). New York: Russell Sage Foundation.

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.

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).

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.

Hill, H., Ball, D. L., & Schilling, S. (2008). Unpacking "pedagogical content knowledge": Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39* (4), 372-400.

Hill, H. C., Rowan, B., Ball, D. L. (2005). Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement. *American Educational Research Journal 42*(2), 371-406.

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

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, 88-11

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.