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.