ABSTRACT

Title of Thesis: AN EXPERIMENTAL STUDY OF MENTORING PRACTICES IN AN AMERICA READS PROGRAM: MEASURES OF INTERVENTION FIDELITY AND IMPLEMENTATION

Janaiha F. Nelson, Master of Arts, 2013

Thesis directed by: Professor Gary D. Gottfredson
Department of Counseling, Higher Education, and Special Education

The America Reads (AR) program at the University of Maryland serves approximately 350 local elementary school students per semester, and trains undergraduate tutors to teach reading using techniques drawn from Reading Recovery methods. Previous research implies that the implementation of interventions should be evaluated prior to gauging their effectiveness. The present study assessed aspects of program implementation for America Reads at the University of Maryland. In addition, it examined the efficacy of a self-monitoring and corrective feedback procedure for improving level of implementation. AR tutors were randomly assigned to the experimental self-monitoring and feedback procedure or to usual and customary monitoring to assess the effects on mentor implementation. Controlling for school assignment, the effect of this self-monitoring and feedback procedure on mentors' self-reported level of implementation was not significant in the small sample of mentors.

Descriptive results including information about the effectiveness and utility of existing procedures for monitoring program implementation, and tutor training have a number of implications for strengthening the Maryland realization of AR; they have implications for the use of monitoring and feedback in the design of similar educational service programs.

AN EXPERIMENTAL STUDY OF MENTORING PRACTICES IN AN AMERICA READS PROGRAM: MEASURES OF INTERVENTION FIDELITY AND IMPLEMENTATION

by

Janaiha F. Nelson

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Arts
2013

Advisory Committee:

Professor Gary D. Gottfredson, Chair
Associate Professor Bill Strein
Affiliate Associate Professor Barbara Jacoby

Acknowledgements

I would like to express my deepest appreciation to advisor and thesis committee chair, Professor Gary Gottfredson. His patient guidance and persistent encouragement are invaluable. I thank Dr. William Strein, and Barbara Jacoby for their support and instruction as thesis committee members. Many thanks to Katie Hershey, and the Team Leaders, Site Supervisors and mentors of the America Reads program who have inspired and supported this research in tangible and intangible ways.

I am thankful to my cohort members who have provided feedback, particularly Ho Lam Yiu and Julie Grossman who provided counsel on previous versions of this paper. Finally, I extend my affectionate gratitude to family, mentors, fellow believers and friends who encouraged me to finish the course.

Table of Contents

List of Tables

List of Abbreviations

AR – America Reads
CTBS – California Test of Basic Skills
DL – Daily logbook
DRA - Developmental Reading Assessment
EOSS – End-of-semester survey
IL – Implementation log
MSS – Mentor satisfaction sheet
OREA – Office of Research Evaluation and Accountability
PGCPS – Prince George’s County Public Schools
RR – Reading Recovery
TL – Team leader
UM – University of Maryland
WWC – What Works Clearinghouse

An Experimental Study of Mentoring Practices in an America Reads Program: Measures of Intervention Fidelity and Implementation

Reading is a foundational skill for learning (Lloyd, 2005). Concerns about reading problems in pre-kindergarten through third grade students have led to a growth in targeted early intervention in the last half-century, with the recognition that early learning experiences and intervention are important for long-term self-efficacy and reading motivation (Fuchs & Morgan, 2007; Gist & Terrell, 1992; Schunk, 1990; Wanzek & Vaughn, 2007; Gambrell, 2011). Among the early intervention systems promoted for reading difficulties is the Reading Recovery (RR) method. Developed in 1976 by Marie Clay in New Zealand, RR is based upon the idea that children learn to read *while* reading (Pinnell, DeFord & Lyons, 1988). Despite the limited and ambiguous data in favor of the efficacy of the Reading Recovery method, the system has spread rapidly in various public school districts through direct advocacy of educators and school administrators (Pinnell, 1990). Elements of the intervention have found their way into numerous early literacy programs (Worthy, Prater, & Pennington, 2003). A discussion of the RR intervention and its efficacy will serve to demonstrate the importance of implementation fidelity both within the RR method and its offshoot literacy interventions, such as the America Reads program.

Reading Recovery Methods

Clay's RR model of reading intervention was brought to the U.S. in 1984, beginning in the Ohio Public School system. The purport of the program is positive effects on children's perceptual analysis of print, knowledge of print conventions, decoding skills, and overall oral language proficiency, in addition to increasing students' prior knowledge, prose comprehension, and strategies related to inference-making and error correction (Pinnell, DeFord & Lyons, 1988).

The overarching goal of the intervention is to close the gap between average and struggling readers. The instructional goal of the method is to promote the development of a “self-extending system” of reading strategies: ostensibly, a system of reading skills that encourages the growth of additional skills (Clay, 1991, p. 317). To do this, the intervention emphasizes scaffolding, building upon students’ existing skills to help them move forward to more challenging concepts. The method uses three main tools to build upon the student’s existing skills: a diagnostic survey, an exploration phase, and an adaptive lesson plan that accounts for the student’s reading level.

The diagnostic procedure includes six assessments: letter identification, high-frequency word recognition, dictation (a test assessing a students’ ability to correctly interpret sounds in words as opposed to correctly spelling words), a writing spree (students are asked to write as many words as possible in an allotted time frame), a text reading (during which the assessor keeps track of frequent mistakes made by the student), and knowledge of print conventions (e.g., where to begin reading on a page) (Clay, 1981; Pinnell, DeFord, & Lyons, 1988). Once the full diagnostic survey has been completed, the next phase of the Reading Recovery (RR) model sets aside time for the student to “explore reading and writing” in order to establish trust and to get a broader sense of the students’ skills (Pinnell, DeFord, & Lyons, 1988, p.16). Finally, the daily intervention involves a 30-minute lesson in which the student has an opportunity to do the following: (1) reread familiar books, (2) read new books, and (3) write and read their own messages and/or stories. During this time, the teacher analyzes the student’s reading using a running record of the student’s reading behavior. The student rereads favorite “familiar” books primarily to increase fluency. As the student reads new books, teachers have the opportunity to teach problem-solving skills to decode unknown words (Pinnell, DeFord, & Lyons, 1988). The student’s progress is monitored carefully through the miscue analysis (or running record).

Several studies have been conducted to investigate the efficacy of the RR method since its inception and propagation in American public schools (Reynolds & Wheldall, 2007). The first of in-house evaluations in the United States was conducted through the Ohio Public School System as a longitudinal pilot study from 1985 to 1988 (Pinnell, DeFord, & Lyons, 1988). Low-performing first-grade readers across 12 sites in Columbus, Ohio were identified for the pilot during the 1985-86 school year (Pinnel, DeFord & Lyons, 1988). The lowest 20% of reading students in the classrooms of Reading Recovery teachers were automatically selected for Reading Recovery. The lowest 20% of reading students of other classrooms were also identified: half were randomly assigned to Reading Recovery and the remainder to a compensatory remedial reading program as comparison (*n* = 51). In total, 136 students were assigned to Reading Recovery. The results provided in the longitudinal study were descriptive, including means from the comparison group of peers in the compensatory program and the group of Reading Recovery students. Reports of RR student means were further divided into (a) students who *successfully* completed the RR curriculum (referred to as "discontinued" students*)* and (b) those who still required remedial support or who moved before completing the program (referred to as "not discontinued" students). At the close of the first year of the study, all the RR students—the aggregate of discontinued and not discontinued students—scored higher than the comparison group on all RR measures, with scores similar to a random sample of typically achieving students. Additionally, RR students on average achieved a gain score of 7.4 points on the National Comprehensive Examination in Reading, in comparison with students in the compensatory program who fell by 2.6 points (Pinnell, DeFord, & Lyons, 1988).

The descriptive means for RR students and compensatory program students implied the program's usefulness for "discontinued" students who successfully completed the program using measures from the diagnostic survey and performance on comprehensive tests of basic skills in reading vocabulary and comprehension (Pinnell, DeFord & Lyons, 1988). However, critics of the longitudinal study pointed out that the aggregation of the data does not sufficiently account for students who did *not* successfully complete the RR program ("not discontinued students). By excluding the Reading Recovery students who were unable to complete the intervention from the aggregation of "discontinued" RR students, the study may have the effect of overestimating the efficacy of intervention (Reynolds & Wheldall, 2007). Apart from these concerns, by earmarking certain students for Reading Recovery, the studied failed to ensure that treatment and comparison groups were randomly equivalent, thus removing its ability to draw any causal inferences from the collected data.

Additionally, it is uncertain how sustained the effects of the intervention are. An examination by Wasik and Slavin (1993) demonstrated that although discontinued students made initial gains, these effects diminished after 12 months and were further reduced at a two-year follow-up. Students for whom the intervention was *not* effective demonstrated the poorest outcomes overall (Pinnell, DeFord, & Lyons, 1988). Approximately 27% students had received RR services, but were unable to discontinue. At the two-year follow-up, these students were reading below average in comparison with both peers who had discontinued from RR and the randomly sampled peers from the population (Reynolds & Wheldall, 2007). The study did not specify the probable impact of regression-selection artifacts that likely affected the groups differently. More specifically, the initial mean reading performance of the lowest twentieth percentile from which the Reading Recovery students were drawn was likely different from the general first grade population with whom they were compared at follow-up.

To bolster the argument that the main effects associated with the intervention were unique to Reading Recovery, an additional in-house Ohio-based efficacy study was conducted in 1994 comparing outcomes for young, at-risk students in RR with those in three other literacy interventions (Pinnell, Lyons, DeFord, Bryk, & Seltzer, 1994). Pinnell et al. examined the effectiveness of Reading Recovery as compared to three other models and their respective comparison groups: Reading Success (a one-on-one tutorial program based on RR), Direct Instruction (a one-on-one skills-based intervention), and a small reading and writing group taught by a trained RR teacher. The unmodified Reading Recovery intervention was the only intervention to demonstrate main effects for standardized and RR measures of reading. The authors concluded that the combination of "individual instruction, instructional emphasis, and teacher professional development" were at the heart of the success of the method (Pinnell et al., 1994, p. 36). On the other hand, a comprehensive review of RR programs found effect sizes as large as .66 in favor of RR; however when other important parameters of instruction were held constant (e.g., teacher training, materials used, etc.), thus isolating "Reading Recovery" specific interventions, the average weighted effect size was .04. The conclusions of the review "did not provide support for the superiority of RR over other one-to-one interventions" (Elbaum et. al, 2000, p. 617).

Other experimental studies examining Reading Recovery interventions present an inconsistent picture of the efficacy of the program. Iverson and Tunmer (1993) compared RR to a modified RR (involving one-to-one explicit instruction in letter-sound patterns) and a standard intervention group. Students in RR groups demonstrated that they were reading at the level of their peers on completion of their program and at the end of the year. The data indicated that students in the modified RR instruction discontinued at a faster rate than traditional RR students, (averaging 41.75 lessons as compared to 57.31 lessons). Baenen, Bernhole, Dulaney, & Banks (1997) conducted a randomized study comparing first grade students in a Reading Recovery intervention condition and a comparison group. For an initial sample size of 168, outcome measures were obtained at three follow-up occasions: the end of first grade (*n* = 147), the end of second grade (*n* = 147), and the end of third grade (*n* = 127). Results indicated that RR students scored significantly higher on the Clay Diagnostic Survey at the first measurement occasion than did the control group. However, by the second year after the close of the study, no significant differences could be found between the groups in test scores, retention rates, or special education referrals.

Schwartz (2005) employed a delayed treatment design to investigate the efficacy of RR. When the initial intervention and delayed intervention groups were measured just before the students in the delayed condition began receiving the RR intervention, it was found that 14% of the delayed intake students had made progress without any intervention. This finding implies that regression artifacts in the gains observed in RR may be associated with students who would have "recovered" toward the mean anyway.

A review of the literature conducted by the What Works Clearinghouse (WWC) (2008) of the U.S. Department of Education summarized the efficacy of the intervention for knowledge of the alphabet ("alphabetics"), fluency, comprehension and general reading achievement. Based on five studies that met its evidence standards for research, the WWC review concluded that the extent of evidence for the Reading Recovery method was medium to large for "alphabetics" and for "general reading achievement," and small for "fluency" and "comprehension. " The WWC review should be considered in light of constraints posed by the available pool of research such as the extent to which randomization was possible, the sample size of the study, adequate study duration, unbiased outcome measures and appropriate weighting for intervention ratings (Slavin, 2008).

Moreover, level of implementation was considered in the evaluative process. Studies indicated that faithful implementation of the system of Reading Recovery was critical to achieving and maintaining effects associated with the program, particularly features such as teacher training, one-on-one attention, and explicit instruction (Elbaum et. al, 2000; Pinnell, Lyons, DeFord, Bryk, & Seltzer, 1994). Whatever efficacy the intervention potentially has could be diluted if it were not faithfully implemented (Fitzgerald, 2001; Worthy, Prater & Pennington, 2003). Issues of treatment fidelity have direct implications for the America Reads program at the University of Maryland, which uses an adaptation of Reading Recovery.

The America Reads Challenge

The America Reads Challenge was a national initiative endorsed during the Clinton administration "to ensure that every child in the United States reads well and independently by the end of third grade" (Roberts, 1999). At present, there are over 1,400 universities nationwide operating America Reads (AR) programs. These programs are only nominally associated with one another and vary widely in the actual interventions they employ. In 1997, AR tutoring commenced at the University of Maryland (UM) in eight elementary schools in the Prince George's County Public Schools (PGCPS) that were selected based on low reading scores, high poverty levels, proximity to university campus, and the availability of a full-time reading specialist at the local school to serve as a site supervisor. Over 80 UM students were recruited and trained as AR "mentors" for the first academic year. The AR program at UM assists first and second grade students who are struggling readers. Students are pulled from class for half an hour, twice a week over a two-and-a-half month period. The program requires mentors to provide at least 14 lessons for each mentee, resulting in a minimum of seven hours of reading instruction per semester. The AR program now sends mentors to twelve public elementary schools in Prince George’s County.

AR mentors are primarily UM undergraduates who participate in one of the following ways: (a) as employees paid through a Federal Work-Study Award, (b) as UM course or internship credit recipients, or (c) as volunteers. Prior to tutoring, mentors receive five hours of training from PGCPS reading specialists. Mentors have the option of requesting two additional training hours in RR techniques, as well as between two and eight hours of training in other tutor-related concerns, including behavior management, effective praise, special considerations for English Language Learners (ELL), community advocacy and education policy. Each mentor is observed at least once during the course of the semester and receives feedback on mentoring practices.

The curriculum for AR at UM was adapted by PGCPS reading specialists from the previously described Reading Recovery methods of Marie Clay (1981). It consists of a 30-minute lesson divided into four sections: student reading, word study, student writing, and read-aloud. In keeping with the RR method, the student-reading portion of the lesson provides opportunities for mentees to reread familiar books to enhance fluency and new books to increase the student’s repertoire of known skills. The writing section is the designated time for mentees to write and read their own sentences and stories. AR omits the Reading Recovery miscue analyses and includes in the thirty-minute lesson a word study section and a read-aloud section. During word study, students and mentors manipulate words with magnetic letters, review flashcards and engage in other drilling exercises. The read-aloud portion is a time for students to listen to mentors’ reading. The AR program at UM retains four of the six formal diagnostic procedures outlined in Marie Clay's method (Pinnell, DeFord, & Lyons, 1988). Pre- and post-assessments on word recognition and spelling, letter recognition, and dictation are completed by AR mentors at the start and conclusion of the two-and-a-half month period. Over the course of the two and a half month period, mentors have an average of about 14 lessons with each mentee, seeing them twice weekly for 30-minute sessions.

Program Efficacy. There have been limited formal evaluations of the America Reads program at the University of Maryland since its launch in 1997. The Prince George's County (Maryland) Public Schools' Office of Research Evaluation and Accountability (OREA) evaluated the America Reads program at the University of Maryland during the 1997 – 1998 academic year (the year following its launch) (Gambrell & Dromsky, 2001). Students from the second-grade at one participating school were randomly assigned to the America Reads program or to a comparison group (*n* = 70) that did not receive tutoring. The sample of 68 intervention-group students was further divided into those who had received 15-45 mentoring sessions (*n = 19*) and 1 – 14 sessions (*n* = 49). On the administration of the California Test of Basic Skills (CTBS) in the spring of 1998, students receiving between 15 and 45 sessions of tutoring scored higher than those students with fewer or no tutoring sessions on measures of Reading Comprehension. This finding approached statistical significance; however, as the sample was drawn from the total school population at the second-grade level it included students customarily not eligible for the America Reads program (i.e., those at or above grade-level in reading). Additionally, as the group of students receiving 15 – 45 was not randomly equivalent to that of students receiving 1-14 sessions, comparisons were biased and may not necessarily provide evidence of efficacy the AR.

In addition to this initial evaluation, annual program data are collected during the course of the academic year, and comparisons of pre- and post- test measures are chronicled, including changes in the book level students were able to read. A book level refers to a system that ranks texts based on reading ease; the higher the book level, the higher the presumed skill of the reader. The America Reads program levels using the Developmental Reading Assessment (DRA). In the academic year preceding the formulation of this study, 2007-2008, 61% of mentees enrolled in the AR program through the University of Maryland moved up three or more book levels and 20% moved up six or more book levels, according to internal office records. Because these are simply pre-/post-assessment comparisons, no direct causal assertions can be made concerning the nature of AR's influence on increases in book level. There limited is quantitative data that describes AR program efficacy, and virtually no quantitative data to describe its implementation. Currently, only qualitative data are available to describe the current level of program *implementation* at each of the mentoring sites. A formal examination of program implementation is a necessary intermediate step in an accurate examination of the efficacy of an intervention (Cook & Shadish, 1986; Sechrest, West, Phillips, Redner & Yeaton, 1979).

Program Implementation and Performance Feedback

In many ways, examining program implementation is the first step in program evaluation as a means of knowing "the specific character of the situation at hand": to examine appropriately the level of implementation requires "scientific fact-finding" (Lewin, 1946, p. 37). Within his action research model (an early form of program evaluation), Lewin outlined a cycle of planning, executing, and reconnaissance that underscores the usefulness of documenting areas of need at each stage of program evaluation. Similarly, Rubin, Stuck, and Revicki (1982) described the importance of establishing a rubric or metric for what constitutes successful implementation within field-based education programs as an essential component of systematic program evaluation.

In some instances, methods employed to examine implementation may in themselves serve as an intervention to improve implementation. In a study originally intended to examine experimental manipulations on worker fatigue, Hawthorne researchers demonstrated that worker performance improved over time during the investigation (Rothlisberger & Dickson, 1930). Originally thought to be due to managerial interest in workers, this result has more recently been interpreted as due to the provision of feedback in conjunction with contingent reinforcement and goal setting (Gottfredson, 2005, p. 2). In their discussion of goal-setting theory, Locke and Latham (2002) described the importance of goal specificity to "direct attention and effort to goal-relevant activities" (p.706). Additionally, they highlighted the need for performance feedback explaining that it is difficult or impossible to adjust various aspects of performance "to what the goal requires" without such feedback (Locke & Latham, 2002, p.708). In summary, the provision of performance feedback, when paired with clarified goals or performance standards, serves to increase performance (Gottfredson, 1996). Setting performance goals or implementation standards and feedback on the degree of achievement of those standards are components of one systematic approach to program evaluation (Gottfredson, 1984).

Treatment Fidelity

Essentially, "what the goal requires" is treatment fidelity. However, strength and integrity of treatments have been cited as a neglected problem in evaluation research (Scott & Sechrest, 1989). A treatment's integrity, or fidelity, refers to whether it has been carried out as planned, and the strength of the treatment conveys the intensity with which it has been implemented. There are many reasons for studying the strength and integrity of a program, the first being its necessity for an adequate examination of program efficacy.

In order to adequately examine a program's efficacy, it must be tested in its strongest form. What on the surface may appear to be an ineffective program may actually be a poorly implemented program. The relevance of treatment integrity is evident in field studies where, in contrast to laboratory studies, conditions are often difficult to standardize (Hulleman & Cordray, 2009). As noted by Sechrest, West, Phillips, Redner and Yeaton (1979), "Real treatments are often complex, are sometimes delivered by poorly trained or unmotivated people, and can be totally disrupted by events in the real world" (p.15). A second important reason for studying fidelity of implementation is that it may provide more specific information about the conditions under which a treatment succeeds or fails with respect to integrity or intensity. Finally, implementation studies may provide information about the feasibility of a proposed treatment (Dusenbury, Brannigan, Falco & Hansen, 2003).

In order to best measure and maximize the integrity of a treatment, the treatment must be well defined. Within the field of public health, fidelity of implementation has been measured in terms of program adherence, dose and quality of program delivery, the engagement of participants in program delivery and the analysis of the program's components (Dusenbury et al., 2003). The criteria for measuring fidelity of implementation are further defined: (a) adherence—whether the components of the intervention are being delivered as designed; (b) duration—the number, length, or frequency of sessions implemented; (c) quality of delivery—the manner in which the implementer delivers the program using the techniques, processes, or methods prescribed; (d) participant responsiveness—the extent to which participants are engaged by and involved in the activities and content of the program; and (e) program differentiation—whether critical features that distinguish the program from the comparison condition are present or absent during implementation.

In order to measure adherence, dose and quality, a formal and specific treatment protocol must be identified. Such a protocol may enhance the likelihood that a treatment is administered in its purest, undiluted form (Scott & Sechrest, 1989). The provision of specific standards for a treatment also increases the ease with which implementation can be measured. The standards, so specified, are better quantified and can be aggregated to provide an estimate of integrity and strength (Hulleman & Cordray, 2009). Mills and Ragan (2000) offer a framework for creating component checklists using a process that identifies the components of the intervention and includes input from past implementers to ascertain ideal use and unacceptable use for each component. The process also involves refining, revising and finalizing the components to be measured and, finally, data collection. There must also be provisions made for departures from indicated standards (Sechrest et al., 1979). In theory (Gottfredson, 1984) better implementation will enhance program efficacy, provided that the theory underlying the intervention is valid and the selected intervention components are known to (or can be shown to) affect causal variables implied by the theory. And in fact, each of the five studies meeting evaluation criteria in a review by O’Donnell (2008) of education-based fidelity studies demonstrated that implementation fidelity is related to program efficacy. The measurement and quality of implementation are, thus, an indispensable aspect of program evaluation.

Despite this clear need for implementation studies tied to program efficacy evaluation for early reading interventions, there are few studies that exist to guide education researchers. Based on her thorough review of existing core curriculum and education research, O’Donnell (2008) offers a set of guidelines for education-based intervention researchers which includes: (a) an a priori identification of the program theory, (b) an operational definition of implementation that specifies essential components, (c) the development of addition measures to examine acceptable adaptations to the intervention, (d) the use of random or full census for generalization purposes, (e) the actual use of the developed measures during the process of implementation and, (f) an examination of reliability and validity of the fidelity data collected (53).

America Reads at the University of Maryland: Measures of Implementation

The America Reads program at the University of Maryland currently uses qualitative observations once during the course of the semester-long program as a means of gauging the level of program implementation and providing performance feedback to mentors. Based on anecdotal observations, inappropriate implementation may take many forms including irregularity of mentoring sessions due to absence or scheduling conflicts, improper time devoted to lesson planning or insufficiently tailored lesson plans, and a host of specific inconsistencies in mentoring technique (e.g., effort-based praise). Although supervisors and program coordinators encourage implementation fidelity, no formal means of monitoring personal performance goals of mentors exist. As a first step toward program evaluation, it is necessary to gather more quantitative data related to program implementation. The present research utilizes an instrument for recording mentors' level of implementation in an experiment designed to test whether self-monitoring procedures can be used to improve the mentors' fidelity to program implementation.

Present Study

The present study seeks to answer both descriptive and causal questions:

1. What is the current level of implementation of the America Reads program?

2. To what extent are self-monitoring of implementation and corrective feedback useful in increasing intervention fidelity?
3. Does prior experience in the AR program (versus no prior experience) interact with self-monitoring in affecting the quality of program implementation?

Method

Participants

The sample consisted of 34 mentors who consented voluntarily to the study from a population of 58 mentors in the America Reads program. Of the volunteers, six subsequently withdrew from further participation in the study, but they permitted the use of data already collected to be used in data analyses. The primary reason cited for study withdrawal was that completing the forms for the experiment was too time-consuming.

Seventy-seven mentors were recruited, selected, and trained for the Spring 2011 semester. Twenty percent of the mentors enrolled for the Spring 2011 semester are in education-related programs of study. Thirty-seven percent are mentoring for first time, 13% are second-time mentors, and 50% have been involved for more than two semesters. Every mentor for the spring 2011 semester was eligible to participate, with the exception of "team leaders", who were involved in observation procedures. Team leaders are mentors that have been promoted to a position of leadership in the program after have mentored for multiple semesters and shown an adequate degree of competency in conducting the intervention. Five participants were male and 29 were female. The participants included 33 undergraduates and one graduate student from diverse of fields of study. Nineteen tutors were first-time mentors, and 15 had prior experience in the program.

Design

The mentors in the study worked in teams at 12 school sites and each school site had one to two team leaders. Participants were randomly assigned within team to one of two conditions, ensuring treatment and control groups balanced by team and equal in expectation on all other variables. The first condition received usual and customary supervision. The second condition consisted of an experimental self-monitoring procedure that provided corrective feedback for level of implementation.

Measures

The following paragraphs describe the measures that were used. A copy of each instrument is located in the appendices.

Team leader observations and implementation log (IL). An implementation log served simultaneously as the treatment and as the form utilized for Team Leader observations. Mentors assigned to the treatment condition used the implementation log to monitor their own performance. Team leaders observed mentors across conditions twice during the study using the implementation log to record mentors' performance. The implementation log is a version of the daily logbook (see Appendix A) that was adapted so that it could be completed using discrete answer choices and its data captured using optical mark recognition. (The daily logbook is described in the auxiliary measures section below.) Mentors received an online instructional video on how to complete the implementation log. The instrument measures the quality and frequency of the intervention—how much mentoring time is received by the mentee, how that time is spent, and whether the 30-minute planning session is utilized. It also captures information about specific elements of the lesson—whether appropriate prompts or an appropriate number of word study aides were used, etc. Each element represented on these measures was given equal weight of one to derive an implementation score, with a maximum score of 20 for each session. The implementation logs were collected weekly. Forms were coded using an ID number to track participant data.

Team leaders assisted in the process of developing a draft of the implementation log to be piloted. Prior to the study, drafts of the implementation logs were given to program team leaders. The team leaders provided informal qualitative feedback on the intelligibility, ease of use, and utility of the implementation log. Appropriate changes were made to the instrument in response to this feedback.

The treatment as administered called for the mentors to provide self-ratings on the Implementation log and to score their own logs for conformity with implementation standards. An examination of the reliability of these self-scored implementation measures found that mentors adequately followed the scoring procedures provided in an instructional video. Each of the implementation measures was independently scored by the investigator and participants' scores were correlated with the score computed by the investigator. Participants' scores clustered around the regression line ($r = .75$; $n = 213$).

End-of-semester survey (EOSS). A post-intervention questionnaire was administered to mentors in both experimental groups to provide self-report information about participants. This self-report instrument reformulated the implementation log to evaluate participants' perception of their work "over the past semester." Participants rated their consistency in implementing various aspects of the mentoring protocol on a scale from 1 to 5, indicating "Never," "Seldom," "About Half the Time," "Usually," and "Almost Always." Two last questions were included to check for potency of treatment manipulations and possible diffusion. More specifically, participants were asked whether they kept personal records of their mentoring sessions and whether they used forms from the other experimental condition.

Prior participation. Length of AR program participation was collected from mentors to gauge the extent of possible interactions of treatment with this characteristic. Prior participation was coded as a dichotomous variable: 0 represented no prior experience and 1, any prior experience.

Auxiliary variables.

Daily logbook (DL). In current usual and customary practice, every mentor indicates what is accomplished during mentoring sessions through self-report. The self-report, called the "daily logbook," is largely open-ended. Mentors record various details of the lesson: how much time was spent on each part of the lesson and general comments about the mentoring session. Participants in both conditions completed the daily logbook for every session of mentoring. Daily logbooks were collected at the close of the semester and were used for informal qualitative analysis of mentoring sessions.

Mentor satisfaction sheet (MSS). Another auxiliary measure was developed in the present study for the purpose of guarding against treatment diffusion and as a manipulation check. Mentors in the control condition completed the mentor satisfaction sheet for the primary purpose of inhibiting treatment diffusion. Assignment to use a specific form was assumed to decrease the likelihood that subjects in the control condition would utilize the "treatment" form. The questions on the MSS related to the availability of supplies, the mentoring environment, and the behavior of the student. Informal qualitative information on the simplicity and utility of the instrument was collected and appropriate changes made prior to the initiation of this study.

Procedure

The experimental calendar provided for nine weeks of data collection. The investigators sought and received approval from the Institutional Review Board of the University of Maryland, College Park. Participants were provided with consent information before receiving AR program training and their school placements. Volunteers who consented were randomly assigned within school to treatment and control conditions. Due to a low sample size, two rounds of informed consent and random assignment were conducted. All participants were stratified by team (i.e., school) prior to random assignment to one of two treatment conditions: (a) control condition and (b) experimental condition. Participants were assigned a computer generated random number. In a software program, participants were sorted by their school team assignment, then by their random number. Alternating participants were assigned to treatment or control groups in the sorted file to minimize imbalance at the school level.

Participants in each condition completed the standard daily logbooks issued by the AR program, submitting them at the close of the semester for informal qualitative analysis. Mentors were given instructions about completion of daily logbooks at initial training sessions, as well as at the site orientations. A supervising team leader observed each mentor twice during the course of the semester, using the implementation log in order to compare level of implementation across conditions. In addition to the daily logbooks that both conditions used, mentors assigned to the treatment condition completed the implementation logs and submitted them weekly. Mentors in the treatment condition calculated their own implementation scores to focus attention on their performance. Mentors did not set personal goals of performance themselves; instead the implementation logs indicated what scores constituted a strong or weak lesson on a scale from 1 to 20. A lesson that scored between 17 to 20 was considered “strong” and was assigned as the performance goal for mentors. A review by Locke and Latham (1990) found that goal setting was a robust intervention and worked irrespective of whether goals were set by participants or assigned, provided that the participants accepted the goals.

Implementation logs and mentor satisfaction sheets were submitted to the AR office every Friday. To encourage prompt submission of implementation logs, study participants received team points (as a part of an ongoing program-wide incentive competition) for turning in their logs weekly. The logs were scanned using optical mark recognition software. Mentors in both conditions were observed twice during the course of the semester using the implementation log as a basis of comparison for level of implementation between treatment and control conditions. At the close of the semester, mentors in both the treatment and control conditions completed an end-of-semester survey to self-report their own daily implementation, as a test for treatment diffusion, and to check the treatment manipulation.

Data Analysis

The dependent variables in this study included the two team-leader observations and the self-reported measure gathered at the close of the study. Both occasions of observation were not completed for every individual participant due to scheduling difficulties—either absence of the mentor or team leader, or absence of a mentee. Moreover, various constraints during the second sampling cycle interfered with the first observation period. As a result, for the first observation period 19 participants were evaluated and for the second observation, 23. Information from the observations was used to describe the level of implementation of AR program components in the University of Maryland at the time of the study.

Data from the End-of-Semester-Survey (EOSS) were used to analyze the second research question, testing the efficacy of self-monitoring of implementation and corrective feedback in increasing intervention fidelity. In total, 25 individuals contributed data for the final self-report to be used in the analysis of treatment efficacy. Data were unavailable for the six participants who withdrew and missing for three additional participants. Table 1 describes the availability of final self-report data.

Table 1
Number of Participants with Self-Report Data by Treatment Condition

	Total	Number with Self-Report Data
Control Condition	16	12 (75.0%)
Treatment Condition (Tx)	18	13 (72.2%)
Total	34	25 (73.5%)

The first research question called for a description of the level of implementation of the AR program. A straightforward tabulation of descriptive statistics on all of the dependent variables was conducted for team leader observations and self-report measures. Although the control group statistics were identified *a priori* as describing the overall level of implementation of the program, tabulations are provided for both the control condition and treatment group in aggregate due to the small sample obtained in the study.

The implementation log included questions pertaining to general lesson information as well as primary components of a lesson plan: (a) student reading, (b) word study, (c) writing, and (d) read aloud (see Appendix B). Descriptive statistics were separately reported for each aspect of the lesson plan as outlined in the implementation log.

The second question called for an assessment of the effects of the experimental intervention on the dependent variables: the level of implementation calculated from the team leader observations on occasions one and two, and the post-intervention self-report measure. It was anticipated that an implementation scale could be derived using the implementation data, and that such a scale might be composed of subscales indicating various model components (e.g., time management, lesson scope, etc.). Given the small sample size, exploratory factor analysis was not feasible. Instead, the overall implementation scale was tested for internal consistency using split-half reliability procedures.

Once reliability tests were conducted, the initial plan was to analyze data using multiple linear regression to conduct significance tests for the treatment effect. The original plan was to examine treatment effects on dependent variables in a two (treatments) by twelve (schools) design. The school-blocking factor is of no particular interest; it is included simply to account for the between-school variance so that the remaining source of error variance is between-subjects-within-schools. The regression model to implement this plan is shown in Equation 1.

$$Y_i = b_0 + b_1 T + \sum_{j=1}^{J-1} b_j S_j + e_i \qquad (1)$$

where Y_i = the implementation score for mentor i, b_0 is the intercept (the mean for control group mentors), b_1 = the treatment effect on implementation, T is a treatment indicator = 1 for mentors in the treatment group and 0 otherwise, b_j = a deviation for school j, S_j is a school indicator variable, and e_i is an error term for mentor i. The ratio of the coefficient for treatment (b_1) to its standard error is a t statistic to be tested for significance with 1 and $(N – J – 2)$ degrees of freedom. Below is a specific description of the indicator coding procedures used.

Table 2

Description of Indicator Coding Procedures for Condition

Experimental Condition	TX_1
Treatment (Implementation Log)	1
Control (Mentor Satisfaction Sheet)	0

Although the approach represented by Equation 1 is to be preferred because it completely removes school as a source of variance, it turned out not to be practical in the present situation due to the small sample size achieved. For this reason, the simpler approach of computing a *t* statistic for the difference between treatment mentor and control mentor implementation measures was applied, and the means and standard deviations for these groups are reported.

Finally, as independent sample *t*-tests are known to be inefficient in small samples, nonparametric tests (the Mann-Whitney *U* Test, the Median test and the Kolmogorov-Smirnov test) were performed and corresponding *p*-values are also reported to provide additional information about differences in this small sample.

The final research question required a test for the interaction of prior experience as a tutor with treatment. The analysis examined the between-subject effects among mentors based on the factors of *experience* and *treatment condition*, using each implementation measure (observations on occasion one and two, and self-report) as the dependent variable, in turn. The dichotomous variable of prior experience grouped mentors into those with "none" (one semester of participation) and "some" (more than one semester of participation) experience. A test for the significance of the increment to the squared multiple correlation due to adding the interaction term was conducted to assess potential statistical interactions.

Results

Current Level of Implementation

Tabulations of descriptive statistics (mean, standard deviation, range, frequency) were calculated across all conditions as a means of establishing the level of implementation for mentors in the program. A comprehensive table of means and standard deviations can be found below (See Table 3). The narrative to follow further explicates the descriptive data according to the categories outlined in the Implementation Log: General Information, Student Reading, Word Study, Writing, and Read-Aloud.

Table 3

Means and Standard Deviations for Implementation on Observed and Self-Reported Occasions

	Observation, First Occasion[a]			Observation, Second Occasion[b]			Mentor Self-Report[c]		
Implementation Item	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*
Total Lesson Length	.89	.32	18	.91	.29	22	4.60	.57	25
Lesson Plan	.68	.47	19	.31	.90	20	3.60	.86	25
Student Reading	1.00	.00	19	1.00	.00	23	4.60	.50	25
Length of Student Reading	.47	.51	19	.68	.47	22	4.20	.64	25
Warm-up Book	.71	.47	17	.91	.28	23	4.54	.83	24
New Book	.94	.23	18	1.00	.00	19	4.40	.86	25
Standard Prompts	.89	.32	18	.68	.47	22	4.56	.65	25
Specific Effort-based Praise	.74	.45	19	.57	.50	21	4.32	.80	25
Teaching Point	.84	.37	19	.65	.48	20	3.76	.92	25
Word Study	.88	.33	17	.96	.20	23	4.68	.62	25
Length of Word Study	.63	.49	19	.74	.44	23	4.24	.77	25
Word Study Tasks	1.00	.00	18	1.00	.00	23	4.60	.64	25
Writing	.93	.25	15	.96	.20	23	4.48	.71	25
Length of Writing	.47	.51	19	.57	.50	23	4.21	.83	24
Student-made Sentence	.74	.45	19	.83	.38	23	4.56	.65	25
Words to Fluency	.58	.50	19	.59	.50	22	4.16	.68	25
Practice Pages	.58	.50	19	.62	.49	21	3.68	1.21	25
Cut-up Sentences	.42	.50	19	.43	.50	21	3.44	1.32	25
Read-Aloud	.46	.51	13	.62	.49	21	3.64	1.11	25
Length of Read-Aloud	.42	.50	19	.55	.51	22	3.92	.95	25

[a] Discrete variable choice (0 = No, 1 = Yes)
[b] Discrete variable choice (0 = No, 1 = Yes)
[c] Likert scale (1=Never, 2=Sometimes, 3=About Half the Time, 4=Almost Always, 5=Always)

General information. Basic information was collected concerning mentors' lesson plans, including the length of the overall session and whether the mentor planned out his or her session prior to executing the session. According to the End-of-Semester Survey, on average mentors reported stayed within the 25 to 35 minute time frame for the lesson "about half the time" to "almost always" (*M*= 4.60, *SD*= .58). This is consistent with Team leader observation.

At the time of observation 1, 89% of mentors' lessons stayed within the acceptable timeframe; 91% of mentors did so at the second observation.

Team leader observations (observation 1 and 2) and mentor self-report indicate a wider range of responses to the question of whether or not mentors planned their lessons prior to working with the student. While on average participants reported that they planned lessons prior to mentoring "usually" or "about half the time" (*M* = 3.60, *SD* = .86), there was a range of endorsed responses for self-report measures, from an endorsed response of "1" indicating the mentor "never" planned lessons in advance to a response "5," indicating mentors "always" planned lessons in advance.

Student Reading. This portion of the lesson provides opportunities for mentees to reread familiar books for fluency and new books to increase unknown skills. Observation measures of participants on both occasions indicated that 100% of mentors conducted a student reading section; End-of-Semester-Survey results are in agreement with the observer data (*M* = 4.60, *SD* = .50). The prescribed length of the student reading (8 to 12 minutes), according to self-report measures was by the mentors was implemented "about half the time" to "almost always" (*M* = 4.20, *SD* = .64). Overall, participants' adherence to the prescribed student reading length increased from 47% at the time of observation 1 (*n* = 19) to 68% at the time of observation 2 (*n* = 22). The descriptive statistics suggest that mentors were more likely to have their student warm up with a familiar book and attempt a new book during the lesson at the time of the second round of observations. Despite this, in the aggregate mentors seemed to use fewer of the specific techniques as outlined in the Reading Recovery (RR) method as the semester continued. In particular, from observation 1 to observation 2, the percentage of mentors using of the standard

RR prompts from the manual, including a teaching point, and employing specific effort-based praised dropped from 89% to 68%, 74% to 57%, and 84% to 65%, respectively.

Word Study. The word study section of the lesson involves activities in which words are manipulated using magnetic letters or flashcards. Letters and words are reviewed as a means of increasing phonemic awareness and sight word recognition, respectively. Mentors self-reported that they consistently included this section of the lesson (*M*= 4.68, *SD*=.62), they kept within the allotted time frame for this section of the lesson (*M*= 4.20, *SD*=.77) and included no more than two activities for this section of the lesson (*M*= 4.60, *SD*=.64). Observation data are in agreement with this depiction. From observation 1 to observation 2, the percentage of mentors including a "Word Study" section increased from 88% to 96% of mentors; adherence to the allotted time frame increased from 63% to 74%.

Writing. Student writing is designated time for mentees to write and read their own sentences and stories. Most participants indicated that they included a writing section in their lesson from "about half the time" to "almost always. Observation measures indicate that mentors regularly included this section (93% at observation 1; 96% at observation 2). Nevertheless, mentors did not adhere as closely to the prescribed Reading Recovery method in terms of specific mechanics. For example, mentors reported using "cut-up sentences" less frequently, with the average mentor doing so "about half the time." Again, this finding was consistent with team leader observations, which found mentors using this particular element less than half of the time (42% - 43%).

Read Aloud. The final portion of the lesson is one in which mentors read to their elementary school student. According to self-report, a wide range of responses was observed for this element of implementation: (responses ranged from "seldom" to "almost always." The average mentor read to their mentee between “about half to time” and “usually” (*M* = 3.64, *SD* = 1.11). At Observation 1, 46% of mentors had included a read aloud; at Observation 2, 62% had done so.

Treatment Effects

Prior to the analysis of treatment effect, the measures were examined for their reliability. Team leader observation measures yielded a corrected split-half reliability of .50, while analysis of the self-report measures demonstrated internal consistency of .84. Although using the team leader observation measure (half of the measure’s variance was due to error) would not bias the findings of the study, doing so would attenuate the effect size of the experimental manipulations, making it more difficult to detect a difference between the treatment and control groups. In contrast, the stepped up split-half reliability for the self-report measure of implementation scale is.84, which is a more adequate level of reliability for group comparisons. The analysis uses the observed measures on occasion one and two, and the End-of-Semester Survey (EOSS) self-report measures as dependent variables (see Table 4).

Table 4
Corrected Split-Half Reliability of Composite Implementation Measures

Measure	*N* cases	*N* items	r_{xx}
Team Leader Observation Measure	23	20	.50
Mentor Self-Report Measure	25	20	.84

Note. Split-half correlations are corrected using the Spearman-Brown formula.

The initial research question called for a statistical analysis that accounted for school team blocks. However, the limited sample size made it necessary to evaluate mentor differences exclusively at the level of treatment assignment. Two models were tested: the first—the one originally proposed—compared treatment and control conditions, controlling for variance at the level of school assignment; the second was simply a significance test at the level of treatment.

Table 5 displays *t*-statistics for the treatment effect on implementation scores from observation occasions one and two, and for the self-report measure, as well as group statistics by treatment and control.

Table 5
Treatment and Control Implementation Score Differences

	Treatment			Control			Difference			
Outcome measure	M_1	SD_1	*n*	M_2	SD_2	*n*	M_1-M_2	*t*	*n*	*p*
Observation (occasion 1)	14.90	3.63	10	15.22	3.03	9	-.32	.21	19	.84
Observation (occasion 2)	15.00	2.38	13	15.60	2.46	10	-.60	.59	23	.56
Mentor self-report	4.31	.33	13	4.10	.40	12	.21	-1.42	25	.17

This finding of no statistically significant effect of treatment is mirrored by a similar result when a vector of school indicators is included in the model to account for clustering within school. Controlling for school assignment, the effect of treatment condition on mentors' self-reported level of implementation is not significant.

As *t*-tests are known to be inefficient in detecting statistical significance in small sample sizes. A series of additional non-parametric tests of the efficacy of the intervention were used to supplement the planned analyses, including the Mann-Whitney *U* test, the median test, and the Kolmogorov-Smirnov test.

The median test examines the null hypothesis that the medians of the populations from which two samples are drawn are the same. The medians of the treatment and control groups on self-reported measures of implementation are found to significantly differ ($\underline{p}$ = .047). The median for the treatment condition was significantly higher than that of the control condition.

The Kolmogorov-Smirov test evaluates the hypothesis that two cumulative distributions could plausibly be drawn from populations with the same cumulative distribution. Here, the difference between the cumulative distributions of treatment and control groups approaches significance (*p* = .069). Again, the treatment condition had a slightly taller cumulative distribution than that of the control condition, indicating better implementation. It is possible that given a larger sample size, significance differences may be found between treatment and control conditions. This possibility would require verification in further research.

The final research question required a test for the interaction of prior experience as a tutor with treatment. The interaction effect of prior experience and the treatment of self-monitoring and feedback on the level of mentor treatment fidelity was nonsignificant. Table 6 represents the *F*-statistics for the interaction model using team leader observations on occasion one and two, and the mentor self-report measure.

Table 6

F-Statistics and Probability Values for Tests for Interaction

Outcome Measure	*F*	*p*	*n*
Observation. First Occasion	.005	.944	19
Observation, Second Occasion	.514	.482	23
Mentor Self-Report	.500	.487	25

Treatment Diffusion

Concerns about possible treatment diffusion prompted the creation of auxiliary variable. The assignment for control group members to use a decoy self-report form was intended to decrease the likelihood that subjects in the control condition would utilize the “treatment” self-report form. The efficacy of this strategy was evaluated through self-report (EOSS). The data suggest that treatment diffusion is not a threat to the internal validity. Of the 25 mentors responding, only one mentor was exposed to the recording document of the other condition. As the mentor in question was assigned to the treatment condition and mistakenly completed a mentor satisfaction sheet rather than the reverse, the experimental manipulation remained intact.

Discussion

The present results apply to the University of Maryland realization of the America Reads program, and they do not necessarily extend beyond the pool of University of Maryland students who participate in the America Reads program and of those, ones who have attended the requisite training and received a school assignment. The results of the study have had (and will have) a variety of implications for implementation monitoring and training of AR mentors at the University of Maryland. They may have similar implications to programs elsewhere with formats similar to that of the Maryland AR program. The study also has a number of implications for planning implementation monitoring systems within programs that rely heavily on a volunteer base of minimally trained and essentially autonomous participants. The discussion to follow addresses the strengths and limitations of the research, examines the study's implications for program enhancement through implementation measurement both within and outside of the UM America Reads program, and suggests directions for future research.

Strengths and Limitations

The experimental design, attention to likely threats to randomization, proactive prevention of attrition, and the development of the observation measure are strengths of the study. Many of the elements incorporated into the design and execution phases could be easily replicated in similar action-oriented program research and evaluation. The observation measure itself was closely anchored to the America Reads program protocol. Once developed, the participation of team leaders was essential in refining the measure. The team leaders evaluated the measure for ease of use and provided suggestions. One such proposal was the creation of a short instructional video on the completion of forms for each condition. Once the measure had been revised and implemented within the study, the measures as completed by participants in the treatment condition were checked against investigator ratings as an assessment of inter-rater reliability. This measure of reliability proved high (*r* = .75, *n* = 213).

Additionally, the execution of the experimental design further demonstrates the feasibility of such a design within an ongoing school-based support program. This design was the best means of drawing confident inferences about the effects of the self-monitoring intervention on the level of implementation achieved by mentors. Furthermore, data collection procedures were put into place to minimize potential treatment diffusion during the study and to check for sources of treatment diffusion at the close of the study through the use of the mentor satisfaction sheet and End-of-Semester Survey (see Appendices).

Despite these virtues, trouble with sampling, participant attrition and missing data made it difficult to make confident causal statements about the effect of immediate self-evaluation and goal feedback on mentors' level and fidelity of implementation. Every mentor recruited and hired for the spring semester was eligible to participate, with the exception of Team Leaders, who conducted observations. Participation was encouraged, but not required. It was anticipated that close to the full pool of available 61 mentors would participate in the study. The initial volunteer sample included less than half that number—29 participants. A second wave of recruitment produced only five additional participants, and there were subsequent withdrawals from the study. The total number of active participants by the close of the study was 25. Had a larger sample size been achieved, the design and analytic approach would have produced a more powerful assessment of the main effects of the treatment as well as the possible interaction of treatment with prior mentoring experience.

Problems with attrition may have been exacerbated by the demands of the study. More specifically, the amount of paperwork involved seems not have been palatable for participants.

Anecdotal self-reports from withdrawing participants indicated that the additional work required in order to complete the implementation log (or mentor satisfaction sheet) was the primary reason they withdrew. Aside from issues with participant and data attrition, the internal consistency of the observation measure was of concern. Low reliability of the observation measure increases the difficulty in discovering group differences. However, while low reliability does present an obstacle, it is a smaller problem in the context of assessing differences between group means than is the small sample size (Stanley, 1971).

Implications for the UM America Reads Program

Upon reviewing the descriptive statistics on implementation, administrators of the UM America Reads program moved forward on four key areas to support increased treatment fidelity: (a) promoting strategies for time-management, (b) providing support for effective lesson-planning, (c) adapting mentor training materials for America Reads “Writing Section” and (d) utilizing implementation logs as a screening measure for new mentees.

Time management. While in general descriptive statistics seem to indicate that mentors’ use of time is within the allotted framework, qualitative review of the team leader observation sheets indicate that mentors often used the maximum time allotted for each section, leaving no time for the Read-Aloud portion of the lesson. To aid mentors in acquiring greater precision in their use of time, administrators implemented two simple interventions. As a first step, administrators distributed a supply of stopwatches in the supply closet, to be used by mentors in their sessions. Further investigation demonstrated that often the Student Reading portion exceeded the time allotted. A simple tool to promote time-management was simply informing mentors that completing a book is not necessary for the effective implementation of the Student Reading section. The student is welcome to read a favorite page from the warm-up activity book, and break up the unfamiliar reading into multiple sections to be completed during different lessons depending on the progress of the students. By making this simple fact more explicit during training sessions, mentors may save time during the Student Reading portion of the lesson. Another strategy that could be implemented to support time management in mentoring sessions also happens to be the second key area targeted for intervention by AR administrators based on the results of the study: that is, lesson-planning.

Lesson-planning. The second occasion of team leader observations indicated that only thirty percent of participants planned lessons in advance of their mentoring sessions. The range of endorsed responses for self-report measures was of concern, confirming that a handful of mentors "never" planned lessons in advance. Yet, planning in advance could support various areas of weak implementation. Mentors who plan in advance would prepare multiple books options in the event of a mismatched difficulty level, organize necessary materials for a smooth writing section, improve in time management and other aspects of implementation. The findings of the study prompted administrators to further the practical supports, if any, being used to foster mentor lesson-planning. Thirty minutes of planning time is scheduled into the mentors' two hours of time on-site; however, administrators found that lack of a specified planning space, tardiness, and the time required to acquire students for mentoring often disrupted or expended the pre-scheduled planning time. In other instances, planning time was used ineffectively as a time for mentors to compare their experiences with other mentors on-site. These discoveries prompted problem-solving as appropriate for each site.

Writing section. Self-report data indicated that mentors included a writing section in their lesson "about half the time" and utilized cut-up sentences less than half of the time (42% - 43%). Further examination found that mentors either did not have adequate time to include a writing-section and its related procedures or they were not sure how to conduct an effective writing section with their student. As a result, the writing section as received greater emphasis in initial mentor trainings. AR trainers have taken the time to include information about the conceptual underpinnings of the writing section rather than purely mechanical to foster a better understanding of its necessity. There is also more time devoted to practicing elements of this section during the training. Additionally, as previously mentioned, strategies for more effective time-management have been explored with mentors as a means of ensuring there is adequate time to complete the writing portion of the lesson.

Implementation log screening. While the study did not conclusively indicate whether or not the implementation log was an effective means of supporting implementation fidelity, team leaders found it to be a useful tool to monitor new mentors and assist them with technique. The implementation logs are now used in the AR program as a screener for new mentees during the first two weeks of mentoring. In this way, team leaders are able to inspect the habits of the mentors prior to the mid-semester observation and provide support early as consistent with the perceived need.

Implications for Program Interventions

Despite its limited power for statistical inferences about the effect of self-monitoring and goal-related feedback, the study demonstrates the practicality of in-house evaluations of implementation. Immediately after results of this research were shared with the administrators of the America Reads program the aforementioned supports were added for areas of perceived weakness in mentor program implementation. The study also has a produced a number of practical considerations for the development of a feasible study of implementation. For example, the number of forms requested for weekly submission may be unnecessary, or the data collection procedures may prove cumbersome. One complication is that the mentors are mostly volunteers, and it is difficult to enforce requirements to maintain records of implementation for volunteers. As indicated in a study of delinquency prevention, difficulty with enforcement is a general weakness of programs that depend on volunteers (Gottfredson et al., 2004). The present study attempted to utilize incentives to *encourage*, rather than enforce record keeping. The incentives did not appear effective, and could be improved and tailored to what volunteers find desirable. One such incentive in the America Reads program could be waving non-essential program requirements. For example, each mentor is required two attend two additional training sessions to remain a mentor in good standing. Training requirements, while beneficial to a mentor's overall knowledge about the field of education (e.g., education policy, community advocacy, etc.), are not necessarily related to program implementation. A more desirable incentive, then, might involve waving one training requirement in exchange for compliance with record keeping.

Future Directions

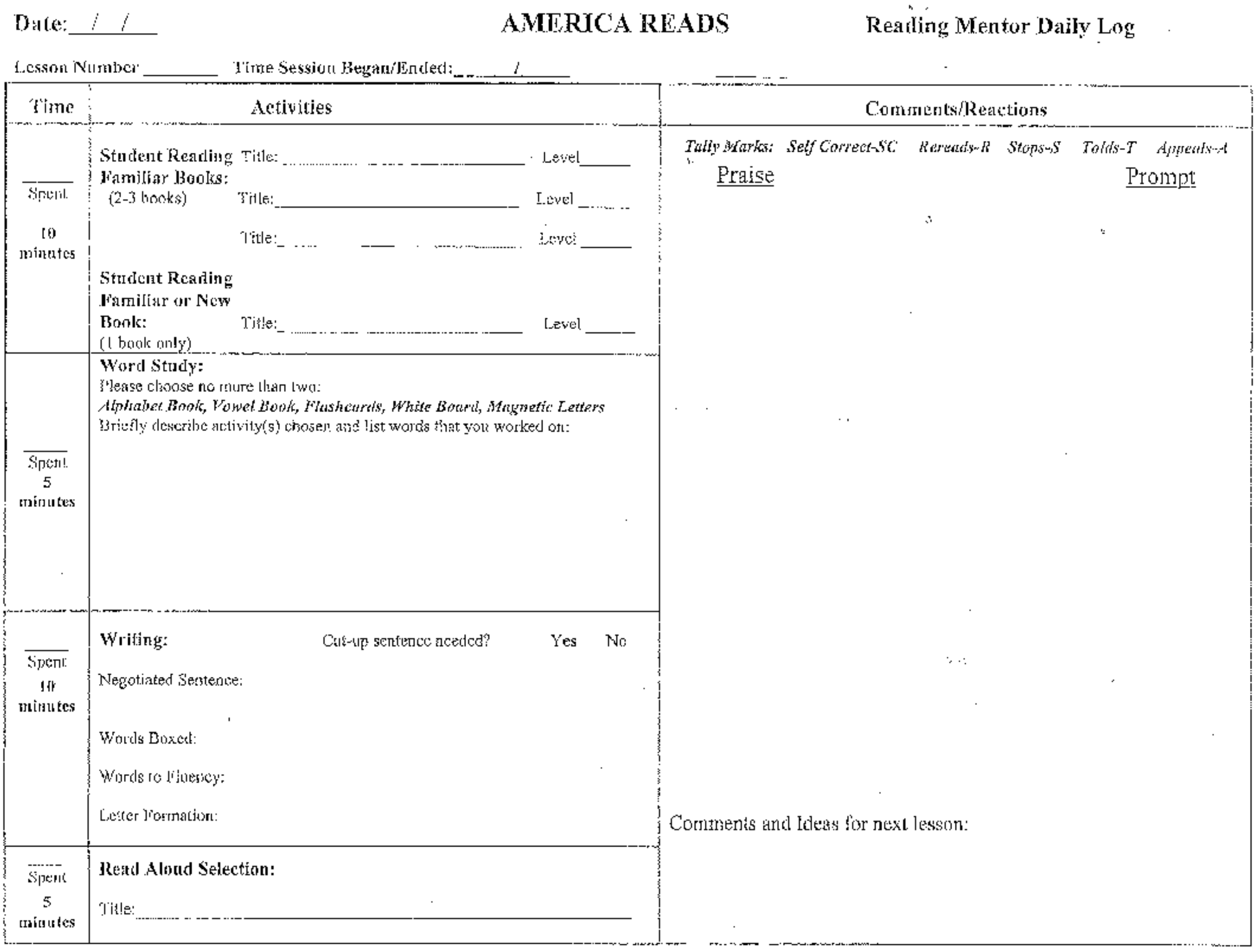
The results of the study may prompt three main categories for future research, including (a) re-evaluations of the efficacy of the implementation log manipulation by adapting a few study parameters, (b) formal survey research investigating treatment fidelity in terms of underlying causes for lack of implementation, and (c) meta-analysis of studies related to early reading interventions with particular attention to feasibility of implementation within a *volunteer* population.

Implementation practices of the America Reads program and experimental manipulations of methods to improve implementation fidelity remain a desirable subject of analytical exploration. It would be useful to re-evaluate the efficacy of self-monitoring and goal feedback procedures as a means of enhancing implementation fidelity, changing the parameters of the study with respect to feasibility and clarity of goal feedback. The first of these changes might involve reducing the number of occasions mentors were required to complete the implementation log. The present study required mentors to complete an implementation log for each occasion of mentoring. For the most active mentor this could entail as many as twelve times per week. In an adapted study, implementation logs could be limited to the first two weeks of the America Reads program, and lifting the requirement to log multiple sessions each day. The reduction in the number of necessary forms may make the study more palatable to participants and increase record keeping compliance. A second potential change would involve dropping the use of Mentor Satisfaction Sheets as a guard against treatment diffusion. Instead, the control condition would use an adapted version of the implementation log that omits the goal feedback information that would theoretically foster implementation fidelity. Finally, in the treatment condition, goal feedback could be made more specific. The current implementation form utilizes a scoring system to indicate a "strong" lesson as opposed to a "weak" one. In addition to quantifying the desirable implementation goals in this way, information concerning non-negotiable elements of the mentor session (i.e., the lesson plan or the cut-up sentence) could be indicated directly on the implementation log. Collectively, these changes would enhance the potential effect produced by the measure, as well as increase the number of forms collected.

The action-oriented nature of the study generated discussion among administrators once data were fed back, and these discussions led to the introduction of several modifications to support areas of weak implementation. Whether these changes ultimately will result in greater treatment fidelity remains to be seen, but the discussions themselves exposed the need for more systematic evaluation of the major causes of poor implementation. More systematic evaluation of the question through formal survey research may assist administrators in targeting program modifications to more precisely address underlying causes. The measures of implementation and treatment fidelity within this study have prompted a wider conversation among program administrations concerning the content of the America Reads program. Using the background knowledge of exploratory implementation studies, content-development could focus on both the efficacy of the interventions being utilized as well as the feasibility of implementation of those interventions.

Appendix A

Daily Logbook



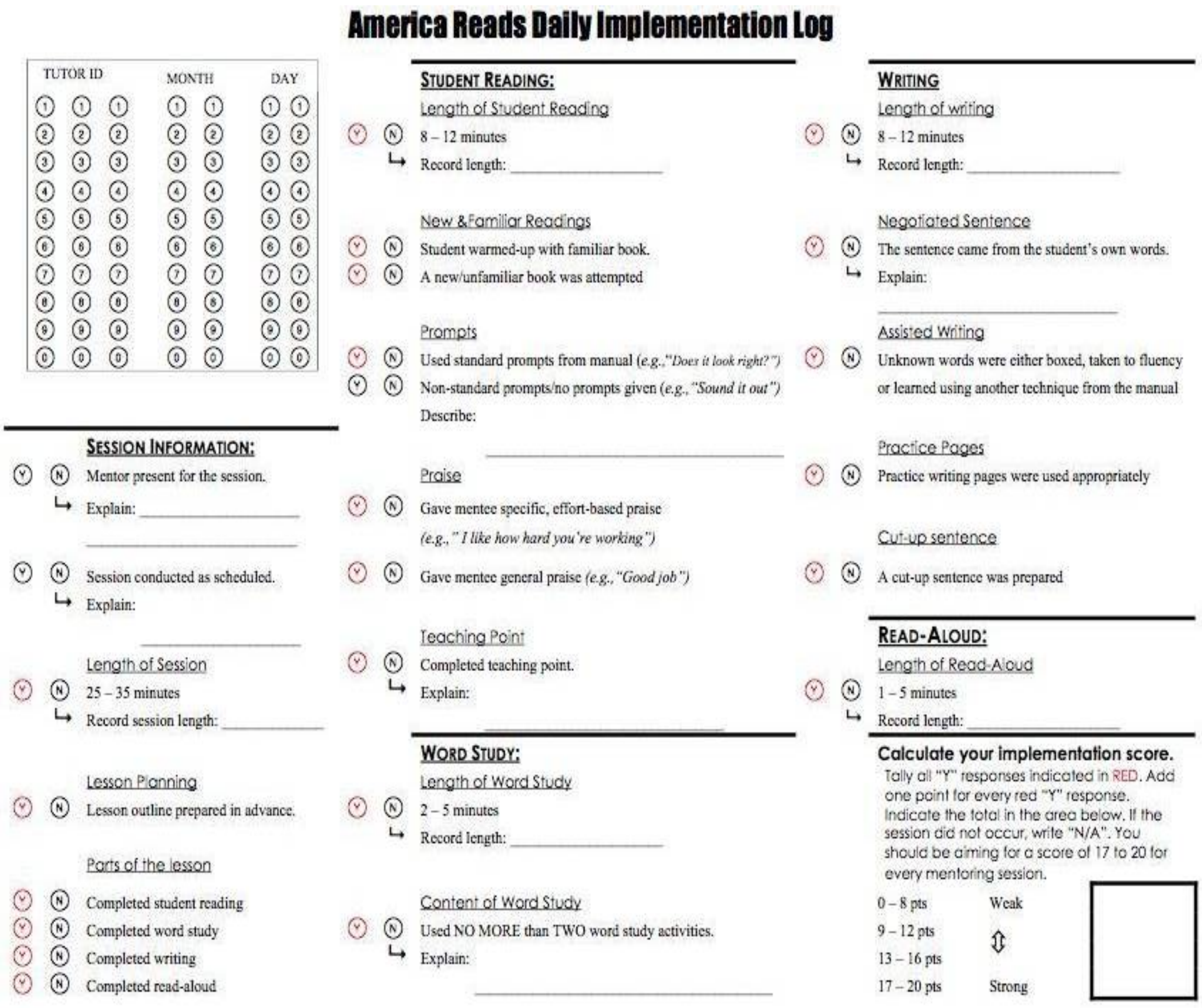
Date:__/__/__ **AMERICA READS** **Reading Mentor Daily Log**

Lesson Number ________ Time Session Began/Ended:______/______

Time	Activities	Comments/Reactions
______ Spent 10 minutes	**Student Reading Familiar Books:** (2-3 books) Title: ________________ Level______ Title:________________ Level ______ Title:________________ Level ______ **Student Reading Familiar or New Book:** (1 book only) Title:________________ Level ______	*Tally Marks: Self Correct-SC Rereads-R Stops-S Tolds-T Appeals-A* Praise Prompt
______ Spent 5 minutes	**Word Study:** Please choose no more than two: ***Alphabet Book, Vowel Book, Flashcards, White Board, Magnetic Letters*** Briefly describe activity(s) chosen and list words that you worked on:	
______ Spent 10 minutes	**Writing:** Cut-up sentence needed? Yes No Negotiated Sentence: Words Boxed: Words to Fluency: Letter Formation:	Comments and Ideas for next lesson:
______ Spent 5 minutes	**Read Aloud Selection:** Title:________________________	

Appendix B

Implementation Log



America Reads Daily Implementation Log

TUTOR ID			MONTH		DAY	
①	①	①	①	①	①	①
②	②	②	②	②	②	②
③	③	③	③	③	③	③
④	④	④	④	④	④	④
⑤	⑤	⑤	⑤	⑤	⑤	⑤
⑥	⑥	⑥	⑥	⑥	⑥	⑥
⑦	⑦	⑦	⑦	⑦	⑦	⑦
⑧	⑧	⑧	⑧	⑧	⑧	⑧
⑨	⑨	⑨	⑨	⑨	⑨	⑨
⓪	⓪	⓪	⓪	⓪	⓪	⓪

SESSION INFORMATION:

Ⓨ Ⓝ Mentor present for the session.
↳ Explain: ____________________

____

Ⓨ Ⓝ Session conducted as scheduled.
↳ Explain:

____

Length of Session

Ⓨ Ⓝ 25 – 35 minutes
↳ Record session length: ____________

Lesson Planning

Ⓨ Ⓝ Lesson outline prepared in advance.

Parts of the lesson

Ⓨ Ⓝ Completed student reading
Ⓨ Ⓝ Completed word study
Ⓨ Ⓝ Completed writing
Ⓨ Ⓝ Completed read-aloud

STUDENT READING:

Length of Student Reading

Ⓨ Ⓝ 8 – 12 minutes
↳ Record length: ____________________

New &Familiar Readings

Ⓨ Ⓝ Student warmed-up with familiar book.
Ⓨ Ⓝ A new/unfamiliar book was attempted

Prompts

Ⓨ Ⓝ Used standard prompts from manual (*e.g.,"Does it look right?"*)
Ⓨ Ⓝ Non-standard prompts/no prompts given (*e.g., "Sound it out"*)
Describe:

____

Praise

Ⓨ Ⓝ Gave mentee specific, effort-based praise
(e.g.," I like how hard you're working")

Ⓨ Ⓝ Gave mentee general praise *(e.g., "Good job")*

Teaching Point

Ⓨ Ⓝ Completed teaching point.
↳ Explain:

____

WORD STUDY:

Length of Word Study

Ⓨ Ⓝ 2 – 5 minutes
↳ Record length: ____________________

Content of Word Study

Ⓨ Ⓝ Used NO MORE than TWO word study activities.
↳ Explain:

____

WRITING

Length of writing

Ⓨ Ⓝ 8 – 12 minutes
↳ Record length: ____________________

Negotiated Sentence

Ⓨ Ⓝ The sentence came from the student's own words.
↳ Explain:

____

Assisted Writing

Ⓨ Ⓝ Unknown words were either boxed, taken to fluency or learned using another technique from the manual

Practice Pages

Ⓨ Ⓝ Practice writing pages were used appropriately

Cut-up sentence

Ⓨ Ⓝ A cut-up sentence was prepared

READ-ALOUD:

Length of Read-Aloud

Ⓨ Ⓝ 1 – 5 minutes
↳ Record length: ____________________

Calculate your implementation score.

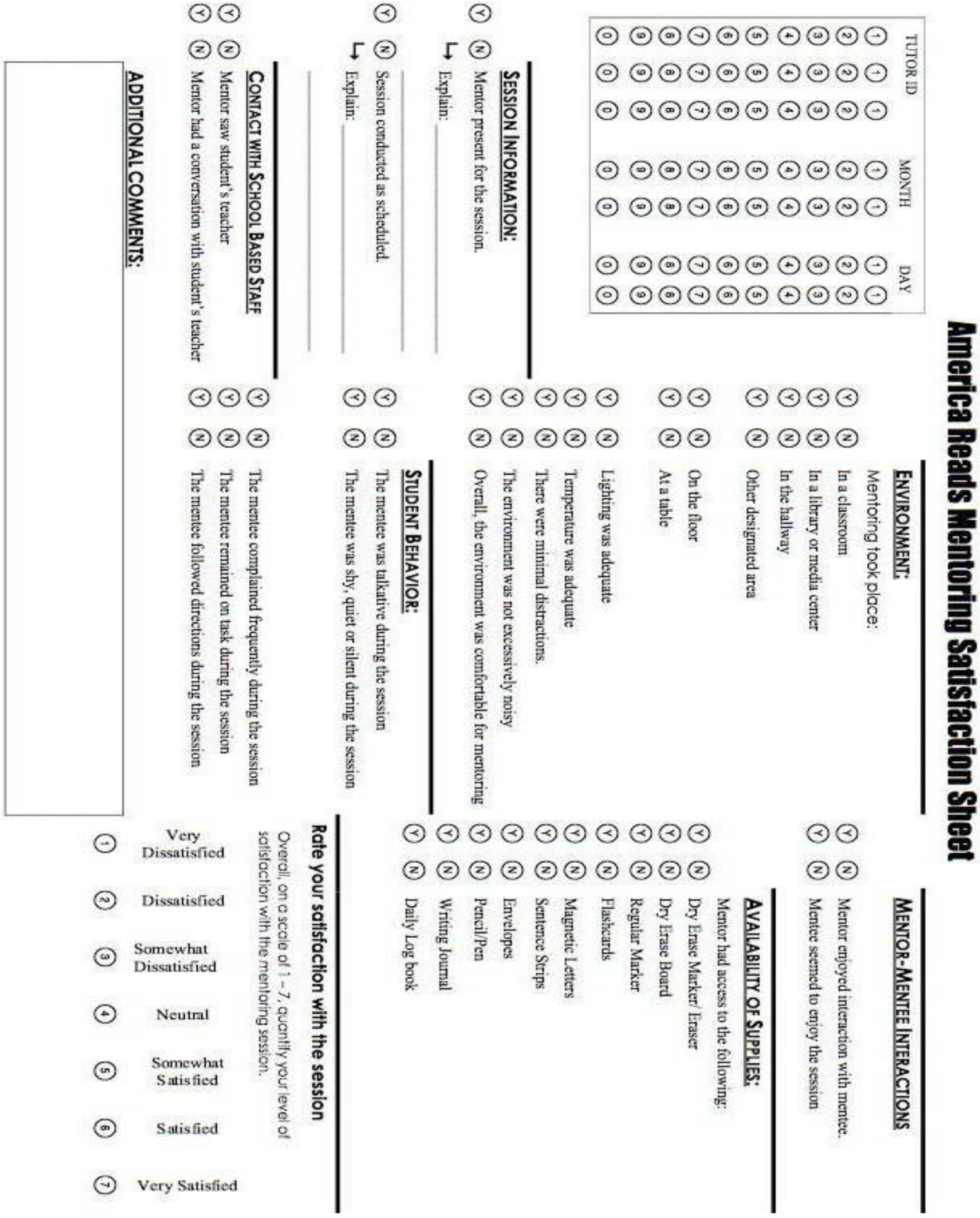
Tally all "Y" responses indicated in RED. Add one point for every red "Y" response. Indicate the total in the area below. If the session did not occur, write "N/A". You should be aiming for a score of 17 to 20 for every mentoring session.

0 – 8 pts	Weak
9 – 12 pts	⇳
13 – 16 pts	
17 – 20 pts	Strong

□

Appendix C

Mentor Satisfaction Sheet



America Reads Mentoring Satisfaction Sheet

TUTOR ID			MONTH		DAY	
①	①	①	①	①	①	①
②	②	②	②	②	②	②
③	③	③	③	③	③	③
④	④	④	④	④	④	④
⑤	⑤	⑤	⑤	⑤	⑤	⑤
⑥	⑥	⑥	⑥	⑥	⑥	⑥
⑦	⑦	⑦	⑦	⑦	⑦	⑦
⑧	⑧	⑧	⑧	⑧	⑧	⑧
⑨	⑨	⑨	⑨	⑨	⑨	⑨
⓪	⓪	⓪	⓪	⓪	⓪	⓪

SESSION INFORMATION:

Ⓨ Ⓝ Mentor present for the session.

↳ Explain: ____________________

____

Ⓨ Ⓝ Session conducted as scheduled.

↳ Explain: ____________________

____

CONTACT WITH SCHOOL BASED STAFF

Ⓨ Ⓝ Mentor saw student's teacher

Ⓨ Ⓝ Mentor had a conversation with student's teacher

ADDITIONAL COMMENTS:

ENVIRONMENT:

Mentoring took place:

Ⓨ Ⓝ In a classroom

Ⓨ Ⓝ In a library or media center

Ⓨ Ⓝ In the hallway

Ⓨ Ⓝ Other designated area

Ⓨ Ⓝ On the floor

Ⓨ Ⓝ At a table

Ⓨ Ⓝ Lighting was adequate

Ⓨ Ⓝ Temperature was adequate

Ⓨ Ⓝ There were minimal distractions.

Ⓨ Ⓝ The environment was not excessively noisy

Ⓨ Ⓝ Overall, the environment was comfortable for mentoring

STUDENT BEHAVIOR:

Ⓨ Ⓝ The mentee was talkative during the session

Ⓨ Ⓝ The mentee was shy, quiet or silent during the session

Ⓨ Ⓝ The mentee complained frequently during the session

Ⓨ Ⓝ The mentee remained on task during the session

Ⓨ Ⓝ The mentee followed directions during the session

MENTOR-MENTEE INTERACTIONS

Ⓨ Ⓝ Mentor enjoyed interaction with mentee.

Ⓨ Ⓝ Mentee seemed to enjoy the session

AVAILABILITY OF SUPPLIES:

Mentor had access to the following:

Ⓨ Ⓝ Dry Erase Marker/ Eraser

Ⓨ Ⓝ Dry Erase Board

Ⓨ Ⓝ Regular Marker

Ⓨ Ⓝ Flashcards

Ⓨ Ⓝ Magnetic Letters

Ⓨ Ⓝ Sentence Strips

Ⓨ Ⓝ Envelopes

Ⓨ Ⓝ Pencil/Pen

Ⓨ Ⓝ Writing Journal

Ⓨ Ⓝ Daily Log book

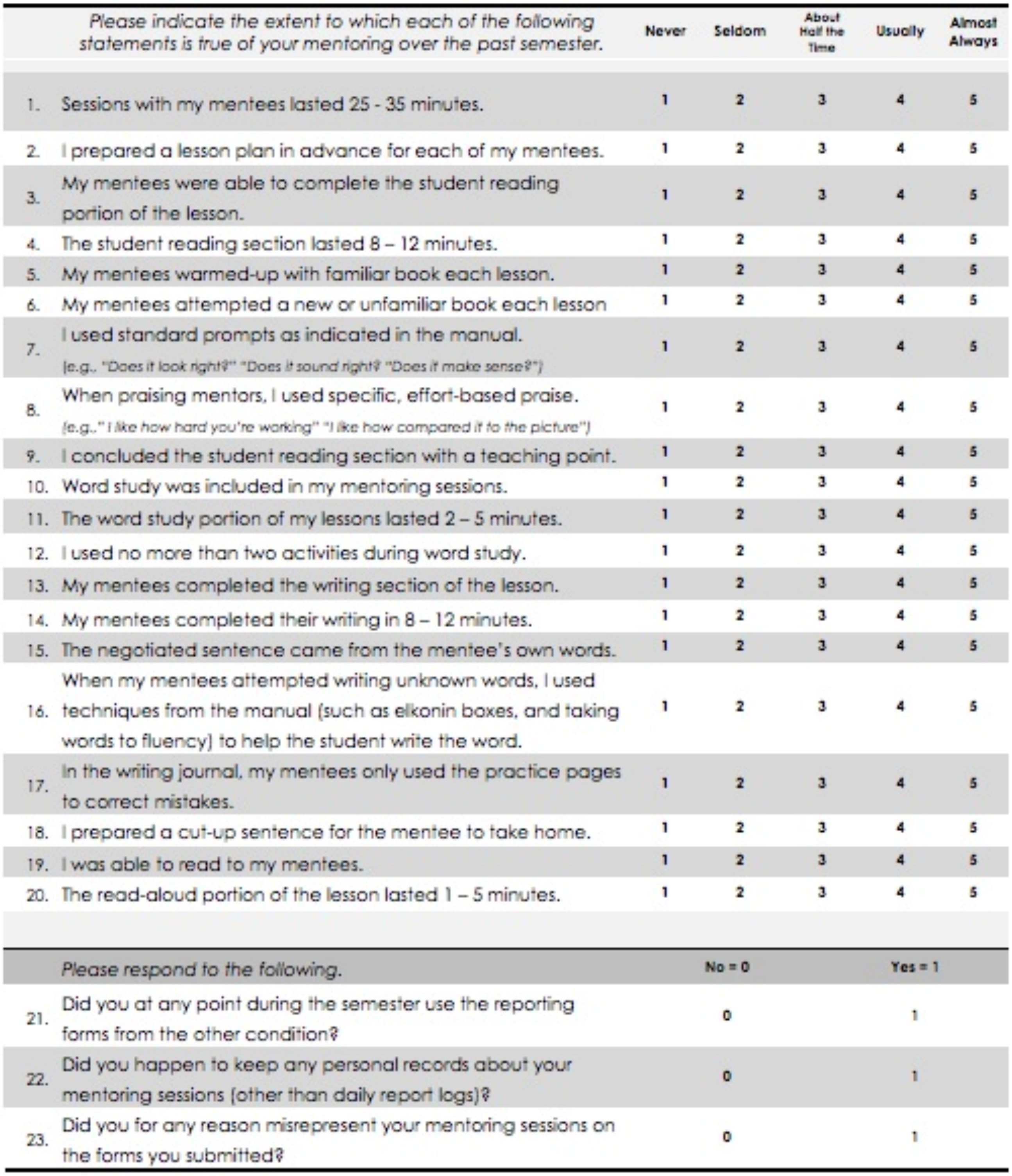
Rate your satisfaction with the session

Overall, on a scale of 1 – 7, quantify your level of satisfaction with the mentoring session.

① Very Dissatisfied	② Dissatisfied	③ Somewhat Dissatisfied	④ Neutral	⑤ Somewhat Satisfied	⑥ Satisfied	⑦ Very Satisfied

Appendix D

End-of-semester survey



	Please indicate the extent to which each of the following statements is true of your mentoring over the past semester.	Never	Seldom	About Half the Time	Usually	Almost Always
1.	Sessions with my mentees lasted 25 - 35 minutes.	1	2	3	4	5
2.	I prepared a lesson plan in advance for each of my mentees.	1	2	3	4	5
3.	My mentees were able to complete the student reading portion of the lesson.	1	2	3	4	5
4.	The student reading section lasted 8 – 12 minutes.	1	2	3	4	5
5.	My mentees warmed-up with familiar book each lesson.	1	2	3	4	5
6.	My mentees attempted a new or unfamiliar book each lesson	1	2	3	4	5
7.	I used standard prompts as indicated in the manual. *(e.g., "Does it look right?" "Does it sound right? "Does it make sense?")*	1	2	3	4	5
8.	When praising mentors, I used specific, effort-based praise. *(e.g.," I like how hard you're working" "I like how compared it to the picture")*	1	2	3	4	5
9.	I concluded the student reading section with a teaching point.	1	2	3	4	5
10.	Word study was included in my mentoring sessions.	1	2	3	4	5
11.	The word study portion of my lessons lasted 2 – 5 minutes.	1	2	3	4	5
12.	I used no more than two activities during word study.	1	2	3	4	5
13.	My mentees completed the writing section of the lesson.	1	2	3	4	5
14.	My mentees completed their writing in 8 – 12 minutes.	1	2	3	4	5
15.	The negotiated sentence came from the mentee's own words.	1	2	3	4	5
16.	When my mentees attempted writing unknown words, I used techniques from the manual (such as elkonin boxes, and taking words to fluency) to help the student write the word.	1	2	3	4	5
17.	In the writing journal, my mentees only used the practice pages to correct mistakes.	1	2	3	4	5
18.	I prepared a cut-up sentence for the mentee to take home.	1	2	3	4	5
19.	I was able to read to my mentees.	1	2	3	4	5
20.	The read-aloud portion of the lesson lasted 1 – 5 minutes.	1	2	3	4	5

	Please respond to the following.	No = 0	Yes = 1
21.	Did you at any point during the semester use the reporting forms from the other condition?	0	1
22.	Did you happen to keep any personal records about your mentoring sessions (other than daily report logs)?	0	1
23.	Did you for any reason misrepresent your mentoring sessions on the forms you submitted?	0	1

References

America Reads*America Counts. (2006). *History*. Retrieved April 29, 2009 from http://www.arac.umd.edu/history.shtml

Baenen, N., Bernhole, A., Dulaney, C., & Banks, K. (1997). Reading Recovery: Long-term progress after three cohorts. *Journal of Education for Students Placed at Risk*, 2, 161-181.

Clay, M. M. (1981). *The early detection of reading difficulties: A diagnostic survey with Reading Recovery procedures.* Portsmouth, NH: Heinemann.

Clay, M. M. (1991). *Becoming literate: The construction of inner control.* Portsmouth, NH: Heinemann.

Cook, T. D., and Shadish, Jr., W. R. (1986). Program evaluation: The worldly science. *Annual Review of Psychology, 37*, 193-232.

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 Theory and Practice, 18,* 237-256.

Elbaum, B., Vaughn, S., Hughes, M. T., & Moody, S. W. (2000). How effective are one-to-one tutoring programs for reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology, 92*, 605–619.

Fitzgerald, J. (2001). Can minimally trained college student volunteers help young at-risk children to read better? *Reading Research Quarterly*, *36*, 28-47.

Gambrell, L.B., & Dromsky, A.J.. (2001). America Reads: Literacy Lessons Learned. In L.M. Morrow, & D.G. Woo. (Eds.), Tutoring Programs for Struggling Readers: The America Reads Challenge (159-173). New Brunswick: Guildford Press.

Gambrell, L.B, (2011). Seven rules of engagement: what's most important to know about motivations to read. *The Reading Teacher, 65*, 172-178. doi:10.1002/TRTR.01024

Gist, M., & Mitchell, T. (1992). Self-efficacy: a theoretical analysis of its determinants and malleability. *Academy of Management Review, 17,* 183-211.

Gottfredson, G. D. (1984). A theory-ridden approach to program evaluation: A method for stimulating researcher-implementer collaboration. *American Psychologist, 39*, 1101-1112.

Gottfredson, G.D. (1996). The Hawthorne misunderstanding (and how to get the Hawthorne effect in action research. *Journal of Research in Crime and Delinquency*, *33*, 28-48.

Gottfredson, G. D. (2005). Hawthorne effect. In B. S. Everitt and D. C. Howell, *Encyclopedia of Statistics in Behavioral Science*, 784-785. Chichester, U.K.: John Wiley & Sons, Ltd.

Gottfredson, G. D., Gottfredson, D. C., Czeh, E. R., Cantor, D., Crosse, S. B., & Hantman, I. (2004). *Toward safe and orderly schools: The National Study of Delinquency Prevention in Schools.* Washington, DC: National Institute of Justice, U.S. Department of Justice.

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

Iverson, S., & Tunmer, W. E. (1993). Phonological processing skills and the Reading Recovery program. *Journal of Educational Psychology, 85*, 112–125.

Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues, 2*, 34-46.

Lloyd, J. W. (2005). *Characteristics of effective reading programs: Promising and not-so promising approaches.* (A summary of the teleconference for the state-to-state Information Sharing Community held on January 25, 2005). Washington, DC: The Access Center.

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting and task performance.* Englewood Cliffs, NJ: Prentice-Hall.

Locke, E. A., & Latham, G.P. (2002). Building a practically useful theory of goal-setting and task motivation: a 35-year Odyssey. *American Psychologist*, *57*, 705-717.

Mills, S. C, & Ragan, T. J. (2000). A tool for analyzing implementation fidelity of an integrated learning system. Educational Technology Research and Development, 48(4), 2-4.

Morgan, P. L., & Fuchs, D. (2007). Is there a bidirectional relationship between children's reading skills and reading motivation? *Exceptional Children*, *73*, 165-183.

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.

Parsons, H. M. (1974). What happened at Hawthorne? *Science, 183*, 922-932.

Pinnell, G. S., DeFord, D. E., & Lyons, C. A. (1988). *Reading Recovery: Early intervention for at-risk first graders*. Arlington, VA: Educational Research Service.

Pinnell, G. S. (1990). Success for low achievers through Reading Recovery. *Educational Leadership*, *48*, 17-21.

Pinnell, G. S., Lyons, C.A., DeFord, D. E., Bryk, A. S., & Seltzer, M. (1994). Comparing instructional models for the literacy education of high-risk first graders. *Reading Research Quarterly*, *29*, 8-39.

Reynolds, M., & Wheldall, K. (2007). Reading Recovery 20 years down the track: Looking forward, looking back. *International Journal of Disability, Development & Education*, *54*, 199-223.

Roberts, A. (November 24, 1999). “The America Reads Challenge.” The Council for Excellence in Government. Retrieved April 28, 2009 from http://www.ed.gov/inits/americareads/aboutus_history.html.

Rothlisberger, F. J., & Dickson, W. J. (1930). *Management and the worker.* Cambridge, MA: Harvard University Press.

Rubin, R., Stuck, G., & Revicki, D. (1982). A model for assessing the degree of implementation in field-based educational programs. *Educational Evaluation and Policy Analysis*, *4*, 189-196.

Sechrest, L., West, S. G., Phillips, M. A., Redner, R., & Yeaton, W. (1979). Some neglected problems in evaluation research: Strength and integrity of treatments. *Evaluation Studies Review Annual, 4*, 15-35.

Schunk, D. H. (1990). Goal setting and self-efficacy during self-regulated learning. *Educational Psychologist, 25*, 71-86.

Schwartz, R. M. (2005). Literacy learning of at-risk first-grade students in the Reading Recovery early intervention. *Journal of Educational Psychology, 97*, 257–267.

Scott, A. G., & Sechrest, L. (1989). Strength of theory and theory of strength. *Evaluation and Program Planning, 14,* 329 – 336.

Slavin, R. (2008). What works? Issues in synthesizing education program evaluations. *Educational Researcher, 37*, 5-14.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (ed.), *Educational measurement* (2nd ed., pp. 356-442). Washington, DC: American Council on Education.

Wasik, B. A., & Slavin, R. E. (1993). Preventing early reading failure with one-to-one tutoring: A review of five programs. *Reading Research Quarterly*, *28*, 179-200.

What Works Clearinghouse (2008). *Reading Recovery (Beginning Reading)***.** U.S. Department of Education**,** Institute of Education Sciences**,** National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. Retrieved August 1, 2012, from http://ies.ed.gov/ncee/wwc/interventionreport.aspx?sid=420

Worthy, J., Prater, K., & Pennington, J. (2003). "It's a program that looks great on paper": The challenge of America Reads. *Journal of Literary Research*, *35*, 879-910.