# Research on Reading Recovery: What is the Impact on Early Literacy Research?

Lea M. McGee
University of Alabama

ABSTRACT

In this article I critique six quantitative studies of Reading Recovery and five reviews of Reading Recovery research published in Tier 1 research journals—journals accepted as having high levels of "expert scrutiny" through peer review. I also critique several quantitative research studies of Reading Recovery published in Tier 2 research journals. These journals are less recognized as outlets for research, or may be perceived of having possible, rather than actual, bias toward more positive views of Reading Recovery. Critique of studies published in Tier 1 research journals revealed many design flaws conducted by researchers aligned with and critical of Reading Recovery, although researchers aligned with Reading Recovery conducted studies with fewer design flaws. The actual findings of five of the six quantitative studies in Tier 1 research journals found positive results for Reading Recovery. Although three studies were critical of Reading Recovery, the results of these studies showed, in two cases, positive outcomes. Several studies including reviews of Reading Recovery reported negative results and these studies were most flawed in design, methodology, statistical analysis, and statement of actual findings. When analyses were examined in detail, the results for Reading Recovery were more positive than reported by the researchers. Research in Tier 2 research journals found positive results for Reading Recovery, and several studies demonstrated longitudinal effects primarily for discontinued students. Overall, Reading Recovery's effect size compares favorably to other large-scale comprehensive school reform models. The research on Reading Recovery provides several insights for early literacy researchers: researchers studying interventions intended to serve the lowest-performing children face many design challenges, educators who read intervention studies must be critical consumers even when studies are published in Tier 1 research journals, and early literacy professionals must be mindful of determining whether children have "deficits" based merely on normal-developing children's performance. I conclude that the positive research outcomes of multiple studies of Reading Recovery, in both Tier 1 and Tier 2 research journals, is a critical finding in early literacy research which must not be undermined by the few negative results found in studies with critical flaws.

Literacy Teaching and Learning
Volume 10, Number 2

pages 1–50

Reading Recovery can trace its roots back 40 years ago when Marie Clay began her careful observation of children as they learned to read and write (Clay, 1966). Since then it has grown into a widely disseminated model of early literacy intervention for struggling first graders with a multi-tiered professional development model for teachers and leaders (Schmitt, Askew, Fountas, Lyons, & Pinnell, 2005). In the United States alone, Reading Recovery has changed the lives of more than 1.6 million children and more than 18,000 teachers. It stands as an unprecedented demonstration of systematic educational redesign that has been successfully replicated in 8,759 schools (during the 2003–2004 school year; Gómez-Bellengé & Thompson, 2005, p. 2). It is impossible to enumerate all the ways in which Reading Recovery has challenged individual teachers, schools, school systems, and the very profession of literacy. The purpose of this review, while situated in this larger picture, is much smaller in scope; I ask, what has been the impact of *research* on Reading Recovery in the field of early literacy research?

First, I need to make clear that I am not a Reading Recovery-trained teacher or leader, and I have never taught at a Reading Recovery University Training Center. So, as I review the research, I am particularly aware that I may make erroneous assumptions regarding Reading Recovery instructional techniques, professional development, underlying theory, or implementation. Nonetheless, it is critical I believe for someone not "inside" Reading Recovery to consider seriously the impact that research on Reading Recovery may have for all early literacy educators and researchers.

Research on Reading Recovery broadly fits into three categories: quantitative research on the effectiveness of Reading Recovery using experimental designs, syntheses of studies examining Reading Recovery effectiveness, and qualitative studies examining teachers' and children's interactions during Reading Recovery lessons. Because of space considerations I will only discuss the first two categories of research: experimental studies and research syntheses. My selection of these two categories does not imply that the third, qualitative research on Reading Recovery teaching and learning, is not critical. Indeed, my reading of this research suggests it is a rich source of insights into the moment-to-moment interactions in which learning occurs. However, I chose to focus this review on the quantitative research because U.S. legislation has validated quantitative research as the primary way to establish effective research-based practice. Reading Recovery has considerable research of this type, conducted by researchers both inside and outside of Reading Recovery; however because of the importance of this type of research, challenges to Reading Recovery's effectiveness based on interpretations of the research have emerged. One of the important outcomes of research on Reading Recovery, as I will show, is understanding the underlying sources of these challenges and what they imply for early literacy.

## QUANTITATIVE RESEARCH ON THE EFFECTIVENESS OF READING RECOVERY

There is a great deal of research which could be considered quantitative research on the effectiveness of Reading Recovery. For example, Reading Recovery's National Data Evaluation Center (NDEC) has for nearly 20 years provided evaluation reports on data for all children enrolled in Reading Recovery and Descubriendo la Lectura in the United States (e.g., Gómez-Bellengé & Thompson, 2005). In addition, there are many reports of the effects of Reading Recovery at both the district and state level, many of which employ research designs using at least quasi-experimental designs (e.g., Briggs & Young, 2003; Brown, Denton, Kelly, & Neal, 1999; Forbes & Szymczuk, 2003; Jaggar & Simic, 1996; Romei, 2002). However, as I began to read this research, I realized that I needed to address issues of "expert scrutiny." That is, my purpose was to review published research; however, it is well recognized that not all research journals subject the manuscripts they publish to the same levels of expert scrutiny during the review process. For example, the Brown et al., (1999) study was published in *ERS Spectrum*, a popular quarterly aimed at school personnel but relatively unknown to academics interested in literacy research. The Forbes and Szymczuk (2003) and Jaggar and Simic (1996) studies were published as technical reports, and the Romei (2002) study remains an unpublished dissertation.

More problematic, for the purposes of my review, was the Briggs and Young (2003) study reported in *The Journal of Reading Recovery*; and even more troublesome were the many other studies published in *Literacy, Teaching and Learning: An International Journal of Early Literacy* (and now titled *Literacy Teaching and Learning: An International Journal of Early Reading and Writing*). *The Journal of Reading Recovery* is published by the Reading Recovery Council of North America (RRCNA). While this journal is a peer-reviewed, refereed journal, the directions to authors clearly state to "select a topic of interest to our Reading Recovery audience" (www.readingrecovery.org/sections/home/newpub2.asp), thus suggesting that manuscripts representing a variety of perspectives—including those critical to Reading Recovery—are not likely to be published in this venue. In contrast, *Literacy Teaching and Learning* is also published by RRCNA and is also peer-reviewed and refereed. However, the journal editors seek submissions related to a variety of topics beyond Reading Recovery that reflect "multiple perspectives and research paradigms from disciplines such as child development, linguistics, literacy education, psychology, public policy, sociology, special education, and teacher education." (www.readingrecovery.org/sections/home/ltl.asp). Many members of the editorial review board are not affiliated with Reading Recovery, thus suggesting that a range of manuscripts including those critical of Reading Recovery may be accepted for publication.

My first decision, therefore, was to determine the kinds of research to include in this review given the issue of differing levels of *perceived* and *actual* expert scrutiny and bias. I decided to do two levels of review: first, review of research published in "Tier 1" research journals that would be expected to have the highest levels of expert scrutiny and second, review of research found in "Tier 2" research journals which may have lower levels of expert scrutiny or may only be perceived as having, rather than actually having, a particular bias. After much reflection, I decided not to include research published in Reading Recovery journals in the review of Tier 1 research journals only because of possible perceived rather than actual bias. I decided not to include technical reports or unpublished dissertations in either level of review.

## REVIEW OF TIER 1 RESEARCH JOURNALS: QUANTITATIVE RESEARCH

In this section I review experimental studies published in top-tier research journals that have investigated the effectiveness of Reading Recovery on children's reading and writing development. Because of length considerations, I do not include studies that have only examined the effect of Reading Recovery on other variables including self-esteem or self-concept (e.g., Cohen, McDonell, & Osborn, 1989; Rumbaugh & Brown, 2000), metacognitive growth (e.g., Cox, Fang, & Schmitt, 1998), or reduction in retention and special education referrals and placements (e.g., Lyons & Beaver, 1995; O'Connor & Simic, 2002). Again, for length considerations I do not discuss investigations of Reading Recovery's cost effectiveness. There have been six investigations of Reading Recovery which have appeared in major research journals. The purpose, sample, design, measures, and outcomes of these studies are summarized in Table 1, beginning on page 6.

Pinnell's (1989) report of two cohorts of Reading Recovery which appeared in *The Elementary School Journal* was the first published research report about the United States' implementation of Reading Recovery. However, Clay (1987) had earlier published research on New Zealand's Reading Recovery achievements in *New Zealand Journal of Educational Studies*, and Shanahan (1987) had published a critique of that research in *Journal of Reading Behavior*. Pinnell's research, although she did not discuss it, is situated within this early interplay between researchers who were publishing research demonstrating the effectiveness of Reading Recovery versus researchers publishing commentary critiquing the research designs or results found in the effectiveness research. For example, that same year Pinnell reported her results, two researchers in New Zealand (Nicholson, 1989; Robinson, 1989) critiqued Clay's (1987) research. These researchers argued that Clay's impressive results must be mitigated by the lack of randomly assigning children to Reading Recovery and other programs,

not using sufficiently sophisticated multivariate statistical methods, and not providing data on school and system change—which she claimed was an important component of Reading Recovery. This interplay between research on Reading Recovery and critiques of it has continued over the past 20 years.

Pinnell (1989) argued that knowledgeable and skillful teachers in New Zealand are less concerned with the polarity between *meaning* approaches versus *code* approaches; instead, they provide a balance of classroom literacy instruction in reading and writing with activities that focus on the details of attending to sound-letter relationships and visual features of print. Thus, she referenced both Goodman (1986) and Chall (1989) as scholars that inform Reading Recovery. She implied that the children who benefit from Reading Recovery were those at-risk children who may not have had the experiences necessary for them to learn from the good instruction they receive, and stated that Reading Recovery does not work with every child. Pinnell credited the success of Reading Recovery to new analytical abilities that teachers use to make decisions about instruction before and during teaching. Finally, she stated that "Reading Recovery is constantly developing" (Pinnell, 1989, p. 180). These four ideas emerged from this research and become contested ground in later research:

1. Reading Recovery does attend to the details of decoding both within reading and writing.

2. Reading Recovery is not the answer for all children but particularly may be useful for children who have not had sufficient prior experiences that allow them to learn from regular classroom instruction.

3. Reading Recovery procedures and methods do change.

4. Reading Recovery changes as a result of close observation and analytical thinking of teachers who have received substantial professional development.

It is important to note that prior to Pinnell's (1989) study, researchers had not addressed the effectiveness of instruction that accelerates learning for the lowest-achieving first-grade readers. This study marks an important transition into research on investigating the effectiveness of interventions rather than on remedial programs for strugglers.

The second major experimental study on the effectiveness of Reading Recovery appeared in 1993 and was published in the *Journal of Educational Psychology* (Iversen & Tunmer, 1993). Details about the design and results of this study are also summarized in Table 1. Iversen and Tunmer chose two previous criticisms of Reading Recovery research to elaborate upon. First, they

**Table 1. Purpose, Sample, Design, Measures, and Results of Reading Recovery Studies Published in Tier 1 Research Journals**

Pinnell, 1989 — *The Elementary School Journal*

| Purpose | Sample | Design |
|---|---|---|
| Provide an introduction to Reading Recovery; describe its theoretical base and its practices, and provide results of 3 years of evaluation studies; most information provided on the 1984 and 1985 years of implementation | 1984 Pilot Study<br>•1984 pilot year<br>•22 RR teachers and 55 RR children<br>•55 comparison children<br>•Schools primarily serve low-income children<br><br>1985 First-Year Implementation<br>•32 RR teachers participated | 1984 Pilot Study<br>RR children from one classroom; comparison children from other randomly selected classrooms. Both RR and comparison were lowest students. RR began in January.<br><br>1985 Implementation<br>RR used lowest-performing children from program classrooms taught by RR teachers. RR and comparison were lowest-performing children from regular classrooms not taught by RR teachers who were randomly assigned to RR program. 102 randomly selected children as "average" comparisons. Note: Numbers of RR and comparison children not included. |

Iversen & Tunmer, 1993 — *Journal of Educational Psychology*

| Purpose | Sample | Design |
|---|---|---|
| Compare Reading Recovery with another "remedial" program delivered to lowest-performing children and to a tutoring program that includes explicit and systematic training in phonological recoding as the only modification of Reading Recovery | •32 RR children<br>•32 modified RR children<br>•32 standard intervention children<br>•(32) classroom controls for RR<br>•(32) classroom controls for modified RR | Complex procedure used to form 3 groups of matched children. Lowest- performing children in schools with modified RR and RR were identified and selected for modified RR or RR. Lowest-performing children in schools without any form of RR were identified and selected for standard intervention. A set of 3 children was formed by selecting a triplet of children in RR, modified RR, and standard classrooms with the same pre-test scores on letter identification and dictation. At discontinuation for each RR and modified RR child, a child from the same classroom identified as "average" was administered measures. Modified RR and RR used same procedures except explicit instruction in letter-phoneme patterns took the place of the letter identification segment of RR lesson. Teachers in modified RR introduced a word, and children built new words by substituting beginning, ending, or medial letters using magnetic letters. |

## Measures

1984 Pilot Study
- 6 measures of Observation Survey
- Stanford Achievement Test

Assessments administered in October, December, and at the end of the year.

1985 Implementation
- 6 measures of Observation Survey
- Writing samples
- 2 subtests of Comprehensive Tests of Basic Skills

## Results

1984 Pilot Study
No inferential statistics; means were similar (or lower) for RR October and December and higher in May: Concepts About Print, Writing Vocabulary, dictation, text level, and SAT.

1985 Implementation
Significant multivariate analysis and univariate t tests; RR children higher than comparison children in both regular (and program) classrooms on 7 out of 9 assessments; 77–90% discontinued RR children performed in average band of performance on 6 Observation Survey subtests.

## Measures

- 6 measures of Observation Survey
- Dolch Word Recognition Test
- Yopp-Singer Phoneme Segmentation Test
- Phoneme Deletion Test
- Pseudoword Decoding Task

Assessments given in beginning of year and at discontinuation for RR, modified RR, and standard intervention children.

Researchers did not specify when the matched standard intervention child was assessed: whether at the time the RR child or the modified RR child was discontinued. The "average" child for each of the 2 RR groups also tested at discontinuation.

Only modified RR and RR children were assessed at the end of the year.

## Results

One-way analyses of variance at pre-test showed no differences. None of the children could respond to any of phonological processing measures.

One-way analyses of variance at discontinuation indicated that all measures were significantly higher for modified RR and RR compared to standard intervention. There were no differences between modified RR and RR except that RR scored higher on phoneme deletion than modified RR.

Matched-pair t tests showed that modified RR and RR children performed no differently than the "average" children and higher than the "average" children on Concepts About Print and phoneme segmentation. RR children also outperformed "average" children on phoneme deletion.

A t-test showed significant differences between number of lessons to discontinuation with modified RR children having fewer lessons compared to RR.

T-tests showed that modified RR and RR children performed no differently at the end of the year except that modified RR children read text at a significantly higher level than RR children. Path analysis showed that end-of-the-year text level was only predicted by end-of-the-year Dolch word reading, that end-of-the-year Dolch reading was only predicted by psuedoword reading at discontinuation, and that psuedoword reading at discontinuation was only predicted by phonological awareness at discontinuation.

**Table 1. CONTINUED**

Pinnell et al., 1994 — *Reading Research Quarterly*

| Purpose | Sample | Design |
|---|---|---|
| Compare RR with other remedial programs to determine whether factors including one-to-one, RR lesson framework, or type and amount of professional development matter in student achievement | 403 children in 10 school districts originally participated<br><br>Pre-test sample: 324 children:<br>•31 RR<br>•41 RS<br>•36 DISP<br>•27 RWG<br>•190 combined control children<br><br>Post-test sample: 283-289 children<br><br>May follow-up sample: 276 children<br><br>October follow-up sample: 274 children | Split plot design replicated over blocks was used. Within a district with RR, 3 other schools were randomly assigned one of three treatment conditions: Reading Success (RS), Direct Instruction Skills Plan (DISP), and Reading/Writing Group (RWG). Each school identified its 10 lowest-scoring children and randomly assigned 4 children to treatment group and the remainder to the comparison group. An additional treatment group of 30 children (Post Study Reading Recovery, PSRR) emerged after post-testing when other children were selected for RR.<br><br>All teachers were videotaped October and May.<br><br>RR instruction was standard Reading Recovery; RS instruction was one-to-one Reading Recovery instruction with teachers having only 2 weeks of professional development; DISP instruction was one-to-one instruction in skills with teachers having 3 days professional development; RWG was small group instruction from RR-trained teachers. |

Center et al., 1995 — *Reading Research Quarterly*

| Purpose | Sample | Design |
|---|---|---|
| Conduct evaluation of RR in which methodological flaws from previous studies have been eliminated: standardized measures, long-term follow-up, random assignment, inclusion of all RR children (discontinued and not discontinued). However, control group was not included in one-to-one instruction, but included in remedial instructional<br>CONTINUES | At pre-test:<br>•31 RR children<br>•39 control children<br>•39 comparison children from matched-comparison schools without RR<br><br>At post-test:<br>•28 RR children<br>•34 control children<br>•36 comparison children<br><br>At short-term maintenance:<br>•22 RR children<br>•31 control children<br>•35 comparison children<br>CONTINUES | In 10 RR schools teachers identified 20 lowest children; from results of Observation Survey, 12 lowest children identified; 8 children were randomly assigned to RR or to control group; other children were to replace RR children when opening occurred so that control children did not go to RR until after short-term follow up.<br>In non-RR schools (matched by region, socioeconomic level, and size) 8 out of 12 lowest-progress children were selected by teacher nomination and randomly assigned to the study.<br><br>Control and comparison children received remedial instruction typically<br>CONTINUES |

## Measures

- Dictation test and text level of Observation Survey (pre-test in October Year 1, post-test in February Year 1, and October Year 2 follow up)
- Mason Early Reading Test (pre-test October Year 1)
- Woodcock Reading Mastery Test-Revised (post-test February Year 1)
- Gates-MacGinitie Reading Test (post-test February Year 1, May follow-up Year 1)

Researchers coded number of minutes in lessons and calculated percentage of time devoted to reading, writing, and other activities.

## Results

Mason and dictation at pre-test were used as covariates in HLM (note, each treatment was compared to its own comparison group randomly selected within same schools). February post-test results were significant for RR in dictation, text level, Woodcock-R, and Gates-MacGinitie; post-tests were significant for RS in dictation and text level; post-tests were significant for RWG for text level. May follow-up tests were significant for PSRR on Gates-MacGinitie. October follow-up tests were significant for RR and PSRR in dictation and text level.

RR lessons were mean of 33 minutes, 60.2% time on reading, 25.3% writing, 14.5% other.

RS lessons were mean of 27 minutes; 62.3% time on reading, 28.8% writing, 8.9% other.

DISP lessons were mean of 27 minutes; 26.8% time on reading, .3% on writing; 69.8% other.

RWG lessons were mean of 32 minutes; 26.8% time reading, 23.4% writing, 49.8% other.

## Measures

- Set 1 comprised of Observation Survey and Burt Word Reading Test
- Set 2 comprised of 6 standardized and criterion-referenced tests: Neale Analysis of Reading Ability–Revised, Passage Reading Test, Waddington Diagnostic Spelling Test, Phonemic Awareness Test, Syntactic Awareness (Cloze) Test, and Word Attack Skills Test

Pre-tests were conducted prior to RR; post-tests were conducted 15 weeks later (at average discontinuation point); short-term maintenance tests were conducted 15 weeks after post-test; medium-term tests were conducted
CONTINUES

## Results

At each of the testing times (pre-test, post-test, short-term maintenance, long-term maintenance), multivariate analysis of variance with repeated measures with alpha set at .05 and univariate pairwise multiple comparisons with alpha set at .01 were conducted. The MANOVA demonstrated an overall significant group by occasion interaction effect.

At pre-test there were no significant differences between the RR and control groups

At post-test (15 weeks after initiation of RR) RR students scored significantly higher than control children on 6 of 8 measures; RR children did not score differently than control children on the Cloze and Phonemic Awareness tests.

At short-term maintenance (15 weeks after post-test) RR students continued to score significantly higher than control
CONTINUES

**Table 1. CONTINUED**

*Center et al., continued*

| Purpose | Sample | Design |
|---|---|---|
| typically available at each school.<br><br>Evaluate contextual, "spillover" systemic effects of RR on educational system by examining low-progress children in schools without RR | At medium-term maintenance:<br>•23 RR children<br>•16 control children<br>•32 comparison children<br><br>15 control children had been selected for and placed in RR after short-term maintenance leaving higher performing, possibly low-normal learners | provided in the schools. Teacher diaries indicated control and comparison children received a variety of remedial assistance for a mean of 2 days per week. |

Chapman, Tunmer, & Prochnow, 2001 — *Scientific Studies of Reading*

| Purpose | Sample | Design |
|---|---|---|
| Examine the relationship between development of phonological processing skills and the effectiveness of RR within a whole language instructional context. Addressed three questions: do RR children in whole language context show deficiencies in phonological processing before RR? Does RR reduce or eliminate deficiencies? Is there a relationship between development of phonological processing skills and immediate and long-term effectiveness of RR? | From a cohort of 152 school entrants in 16 schools:<br>•26 RR children who were discontinued<br>•6 RR children who were referred<br>•20 controls<br>•80 normally-developing children | Children were identified at beginning of Year 2 for RR using Observation Survey. 32 children from 16 schools were selected. A control group was selected using Year 1 end-of-the-year scores: 20 children "whose end of the year mean scores on context-free word identification and book level were similar to those of the RR children" were selected to form the control group. By the middle of Year 2, 6 RR children were referred on and 26 RR children were discontinued forming two RR groups (only RR discontinued children were included in analyses). 80 normal-developing children were selected so their scores were higher than RR and control. Note, out of the 152 children, 20 children were not included in the study. It is not clear whether RR and control children are in the same schools or classrooms, nor what "similar" performance indicates. It is not clear why only 20 control children were selected rather than 32 or 26. Thus, individual children were not matched. |

| Measures | Results |
|---|---|
| 12 months after post-test (one year after average discontinuation point).<br><br>RR and control children were tested on Set 1 and Set 2 assessments at pre-test and post-test; at short-term maintenance RR and control children were tested on Set 2, Burt Word, and book level only; at medium-term maintenance control and RR children were tested on Set 2, Burt Word, book level, and passage comprehension from Woodcock Reading Mastery Test. Comparison children received Set 2 at pre-test, post-test, short-term maintenance, and medium-term maintenance; they were also tested on Burt Word, book level, and passage comprehension test from Woodcock Reading Mastery Test at medium-term maintenance. | children on all measures except the Cloze and Word Attack Skills tests. However, the Word Attack Skills Test was nearly significant (p. 011). MANOVA analyses revealed no differences between control group and comparison group children at pre-test, post-test, and short-term maintenance.<br><br>At medium-term maintenance (12 months after post-test), the researchers reported that MANOVA demonstrated no differences between RR and control group, although the reported p value was significant [F (8,30) = .262, p. = 0268), p.253]. The researchers reported that follow-up univariate analyses revealed only one significant difference at the p. 01 value: book level. No statistical comparison of control and comparison group children was undertaken at this time occasion. |

| Measures | Results |
|---|---|
| Children were assessed 6 times: beginning, middle, end of Year 1; middle and end of Year 2; end of Year 3.<br><br>Assessments were phoneme deletion, sound matching task, phoneme segmentation task, letter-name, verbal working memory, receptive vocabulary, pseudoword decoding, analogical transfer task, contextual facilitation task, context-free word recognition ability, Burt Word Reading Test, Neale Analysis of Reading Ability–Revised (accuracy subtest), preconventional spelling, conventional spelling, Interactive Reading Assessment System (reading comprehension subtest), book level, reading self-concept, academic self-concept, adaptive functioning and behavioral problems.<br><br>Two groups of RR children were identified based on Burt Word scores: 22 minimal benefits children (1-13 on Burt), 7 moderate benefits group (16-26 Burt). | At Time 1, analyses of variance with follow-up Scheffe tests demonstrated that RR differed from ND on most measures except verbal working memory and reading concept. Control differed from ND on most measures except phoneme deletion, verbal working memory, and reading self-concept. RR and control did not differ on any measure.<br><br>At Time 2, analysis of variance with follow-up Scheffe tests showed that RR differed from ND on all measures except verbal working memory whereas control differed from ND on phoneme deletion and verbal working memory. Control and RR did not differ on any measure.<br><br>At Time 3 RR and control differed from ND on all measures except reading self-concept and only differed from each other on preconventional spelling with control scoring higher than RR.<br><br>At Times 4, 5, and 6, analysis of variance with follow-up Scheffe tests showed that RR and control differed from ND on all measures except phoneme segmentation at Time 5 and 6. RR and control did not differ except at Time 4 and 6 on analogical transfer. RR and PRC differed from ND children on reading and academic self-concept at Time 6.<br><br>T-tests on phonological measures at Time 3 and 4 for minimal and moderate benefits group were significant. |

**Table 1. CONTINUED**

Schwartz, 2005 — *Journal of Educational Psychology*

| Purpose | Sample | Design |
|---|---|---|
| Investigate (a) whether RR increases performance compared to similar students without RR, (b) whether RR closes the gap with average-performing children, (c) percentage of children identified as needing RR who actually make adequate gains without RR, and (d) percentage of children who need long-term support in literacy after RR.<br><br>To eliminate other design flaws: classroom/teacher effects and random assignment | •36 RR first-round children<br>•36 second-round children<br>•29 low-average children<br>•32 high-average children | 37 RR teachers selected lowest-performing children and selected 3 of lowest to form 3 slots in their teaching. The 2 next lowest-performing children were randomly assigned either to RR at the beginning of the year or to second-round service. To control for classroom/teacher effects, RR children and control children (selected for second-round service) were in the same classroom. In addition, from the same classroom the RR teachers selected a low-average and a high-average child based on teacher rankings.<br><br>Matched pairs of first- and second-round children in same classroom were established to determine efficiency, establishing the percentage of children who would have succeeded without RR. At transition first-round RR children were expected to reach text levels at or above 12 and second-round students were expected to read at or below text level 6. |

## Measures

- 6 measures of the Observation Survey were administered pre-treatment, transition, and end of the year

- Yopp-Singer Phoneme Segmentation Task
- Sound Deletion task (10-item version of Rosner task)
- Slosson Oral Reading Test–Revised
- Degrees of Reading Power Test

were all administered at transition and end of the year

## Results

Repeated analyses of variance at 3 test times (pre-treatment, transition, end of the year) showed significant group x test period interactions for each Observation Survey subtest. Repeated analysis of variance at 2 test times (transition and end of the year) showed significant interactions between group and test time for Degrees of Reading Power and phoneme segmentation. Slosson Oral Reading Text-Revised demonstrated main effect for group and the phoneme deletion measure had a main effect for time. Simple effects and comparisons at pre-treatment showed significant group differences with first- and second-round RR children scoring lower than high-average on all measures and lower than the low-average children on Ohio Word Test, Writing Vocabulary, Hearing and Recording Sounds in Words. There were no differences between first- and second-round RR children. Simple effects and comparisons at transition first-round RR children scored significantly higher than second-round RR children on all Observation Survey subtests (effect sizes ranged from .90-2.02). First-round RR children scored higher than second-round children on Slosson (d=.94). There were no significant differences between high-average and first-round RR children. First-round RR children scored higher than low-average children on text level, and marginally higher on Ohio Word Test, and Concepts About Print. Simple effects and comparisons at end of the year showed no differences between the first-round RR children and the other 3 groups (second-round RR children, low-average, high-average). RR children scored marginally lower than high-average on text level and Degrees of Reading Power. The second-round RR group scored lower than the low- and high-average group on text level.

62% matched pairs confirmed the expected pattern.

24% matched pairs showed that both first-round and second-round RR children read below expected text levels.

11% matched pairs showed that while first-round RR children read at or above expected levels, second-round children also read above expected.

3% matched pairs showed that first-round RR children did not meet text level expectations, whereas second-round RR children read above expected.

Thus, 86% of second-round children made little progress without RR whereas 14% made better than expected progress. In contrast, 27% of the first-round RR students did not meet text level expectations.

argued from previous evaluations (Glynn, Crooks, Bethune, Ballard, & Smith, 1989) that the text level measure in *An Observation Survey of Early Literacy Achievement* (Clay, 2006) does not provide interval data (children need less time to make a gain of one level of text at low text levels compared to high levels). [Note: because the assessment measures included in the Observation Survey have not changed substantially, throughout the manuscript I refer to the revised second edition published in 2006, although researchers used earlier available editions.] Further, they argued that previous research had only compared one-to-one Reading Recovery to group-administered remedial programs. However, much of their rationale focused on the need for "explicit and systematic training in phonological recoding skills" (Iversen & Tunmer, 1993, p. 113) which they argued that Clay suggested was not necessary. Instead, their interpretation of Clay was that instruction in the alphabetic code and using phonological recoding in Reading Recovery would arise within the context of reading and writing without resort to explicit instruction in the use of letter-to-phoneme correspondences. Moreover, Iversen and Tunmer suggested that using spelling to teach sound-letter correspondences, as they argued is the case in Reading Recovery, would not work because "Thompson, Fletcher-Finn, and Cottrell (1991) found that knowledge of phoneme-to-letter correspondences acquired through spelling did not automatically transfer as a source of knowledge for letter-to-phoneme correspondences in reading" (Iversen & Tunmer, 1993, p. 114). Yet, their argument that letter-sound instruction during spelling is ineffective is based on this one unpublished manuscript.

There are several other concerns to note about this study. First, the authors (one connected to Reading Recovery and one not) did not resolve several of the methodological problems previously noted by critics. That is, they did not randomly assign children to treatment and control groups; their major control group was a small-group intervention rather than one-to-one instruction although they did also use a one-to-one modified Reading Recovery group as a control (the modification included a few minutes of direct instruction in making and breaking words). Further, they did not use multivariate statistics. Finally, the authors overstated their results. For example, they answered their question, "Is Clay (1985, 1991) correct in arguing that instruction in alphabetic coding should normally arise incidentally in the context of reading connected text and that explicit instruction in the use of letter-to-phoneme correspondences is largely unnecessary because children can acquire knowledge of the alphabetic code through their experiences of learning to spell words?" (Iversen & Tunmer, 1993, p. 123), with a clear no. However, Reading Recovery children scored better than average control children in the very variables of interest; they were better able to segment and delete phonemes whereas the modified Reading Recovery children were only better at segmenting. However, the modified Reading Recovery program was able to reach these results in

15 fewer lessons (41.75 versus 57.31 lessons). Thus, the researchers suggested that some more explicit attention to letter-sound relationships within the Reading Recovery lesson framework may increase cost effectiveness.

There are at least two reasons to be cautious about interpreting this finding. The major purpose of this study was to demonstrate that two different versions of Reading Recovery were effective compared to small-group instruction even when the research was conducted by at least one researcher openly critical of Reading Recovery. The analysis of the number of lessons until discontinuation was not planned, but emerged in the study suggesting that replication is needed to confirm this finding. A more valid method of establishing a difference in time of discontinuation for children who have more explicit attention to letter-sound relationships within Reading Recovery is to assess both groups after a specified number of lessons (for example, 30 or 40) to examine differences in performance. End of program decisions are somewhat subjective and in this case made by the first author of the study.

It is important to note that while the modified Reading Recovery lessons included more minutes of instruction on working directly with words, the instruction did not follow a script nor predetermined sequence. Instead, modified Reading Recovery teachers selected words (not single letters and sounds) from books children read in which to break and build new words focusing on beginning, ending, and medial sounds. Thus, the authors' claim that "systematic instruction in phonological recoding skills was more effective than incidental instruction" (Iversen & Tunmer, 1992, p. 123) overstates their findings; teachers did not employ systematic instruction. Instead they used direct word-building-and-breaking instruction and found that direct instruction in making and breaking words as a portion of Reading Recovery led to faster discontinuation rates but not to greater phonological skills, text reading abilities, or writing abilities at the point of discontinuation. Researchers not familiar with Reading Recovery would not know that this making and breaking word practice was added to Reading Recovery techniques in 1993 (Clay, 1993) independent of this research study.

In 1994, Pinnell, Lyons, DeFord, Bryk, and Seltzer published the results of an extensive study of the effectiveness of specific components of Reading Recovery (summarized in Table 1). They examined whether it was the one-to-one instruction, professional development, or the massive amounts of reading and writing which emerge from the Reading Recovery lesson framework that could account for accelerated learning. This study investigated the various "values added" that Reading Recovery delivers to children beyond regular classroom instruction. The first value added to regular classroom instruction is one-to-one instruction. Thus, Reading Recovery (RR) was compared to a reading and writing group (RWG) in which teachers had access to the same materials, had received the same extensive professional development, but taught

a small group of children rather than one child. The second value added is a yearlong extensive professional development that involves teaching before peers (behind-the-glass teaching) and reflection to build a strong consistent theoretical orientation that leads to teacher problem solving at the point of difficulty in teaching. Thus, Reading Recovery was compared to Reading Success (RS) in which teachers received only 2 weeks of professional development, still used one-to-one instruction, and engaged children in massive amounts of reading and writing using the Reading Recovery lesson framework. The third value added is the lesson framework which maximizes the amount of time children spend reading and writing. Thus, Reading Recovery was compared to another intervention program, Direct Instruction Skills Plan (DISP) in which teachers provided one-to-one instruction, but did not use Reading Recovery materials or the lesson framework. They also did not have extensive professional development in delivering their instructional program.

Later critique of the Pinnell et al. (1994) study revealed a "value subtracted" component in the study. Rasinski (1995) argued that the results could be attributed to the greater skills and experience of the RR and RWG teachers compared to the DISP and RS teachers. These later two groups of teachers were substitute teachers whereas the former two groups of teachers were employed in the school system and experienced in their use of Reading Recovery. Thus, the added values of Reading Recovery beyond regular classroom instruction include one-to-one instruction, extensive professional development, and massive amounts of reading and writing. The values of the RWG included extensive professional development. The values of RS were one-to-one instruction and massive amount of reading and writing, but subtracted from that is experience in teaching. The value of the DISP was one-to-one instruction, but subtracted is experience in teaching.

The results of this study suggest that some of these values are more critical in helping children succeed in learning to read than others. First, one-to-one instruction without massive amounts of reading and writing, professional development, and experience produces no effects beyond regular classroom instruction (see results of DISP). Second, extensive professional development and experience without extensive amounts of reading and writing or one-to-one instruction provides only a small effect beyond regular classroom instruction (see results of RWG). Third, a lesson framework including extensive amounts of reading and writing with one-to-one support can produce some improvements beyond regular classroom instruction even without extensive professional development (see results of RS). However, the extensive experience and professional development of Reading Recovery teachers in one-to-one instruction produces results beyond those obtained merely using the lesson framework or materials even in one-to-one settings.

These Reading Recovery researchers addressed each of the design criticisms of Reading Recovery research to date. They used a complex research design

with random assignment of treatments to schools, random assignment of children to treatment or control, and sophisticated multivariate statistics (HLM) using covariants. Control and treatment children were from the same school (although not necessarily the same classrooms). Unfortunately, the design called for a control group specific to each treatment group and, thus, did not allow comparison among the treatment groups. However, effect sizes were computed. Moreover, the researchers videotaped in each classroom and carefully observed two lessons in order to calculate the number of minutes of instruction and the type of activities involved in lessons. Although the authors did not discuss fidelity to treatment, their rich description of the lesson activities suggested that teachers did conform to their treatment. They included standardized reading measures beyond the Observation Survey and compared the one-to-one Reading Recovery intervention treatment to another one-to-one treatment rather than merely to small-group remediation already in place. All Reading Recovery children were included in the analyses regardless of whether they had discontinued, providing a more comprehensive picture of the true value added by Reading Recovery. The inclusion of all children who participated in Reading Recovery instruction may have contributed to the small to moderate effect sizes reported in this study.

An unexpected finding in this study was that Reading Recovery-trained teachers did not engage children in extensive amounts of reading and writing during small-group instruction (RWG). Massive amounts of reading and writing are hallmarks of Reading Recovery lessons and even less well-trained and inexperienced teachers were able to accomplish this during RS lessons. Therefore, the difference in amount of reading and writing time in RWG may be related to the small group versus one-to-one instructional context rather than to amount of training. The authors called for additional work in designing and research on the effects of small-group lessons that maximize time spent reading and writing in small-group instruction.

The fourth major quantitative study of Reading Recovery appeared in 1995. Center, Wheldall, Freeman, Outhred, and McNaught examined Reading Recovery's effectiveness using a design that eliminated some of the earlier weaknesses, although control and comparison children did not receive one-to-one instruction (summarized in Table 1). Reading Recovery was found to produce superior results at the point at which most children would discontinue, 15 weeks later (short-term maintenance), but not 1 year after discontinuation (medium-term maintenance). However, these results could be due to the loss of 15 out of 31 children in the control group who were considered so low that they had entered Reading Recovery at the time of medium-term maintenance. The 16 children remaining in the control group were the highest-performing children in this group.

There are two reasons to suspect that this attrition was the cause of the results rather than actual waning of Reading Recovery's effectiveness. At

discontinuation and at short-term maintenance (15 weeks after discontinuation) the performance of the control group and comparison groups (low-performing children in schools that did not have Reading Recovery) were found to be the same. These two groups were not compared at medium-term maintenance (1 year later) because "inspection of the means of the control and comparison groups at this testing period indicates only a slight superiority on the part of the control group. Since many of the weaker students in this group had been removed, it is reasonable to assume that the differences between the two groups at the medium-term evaluation would have been marginal and unlikely to have been statistically significant" (Center et al., 1995, p. 260). Note that the authors did not statistically compare control with the comparison group at this point "because of selective attenuation of the control group" (p. 260). Yet, they did statistically compare this same control group to Reading Recovery children and found that the MANOVA only approached significance. My inspection of the means of the comparison group suggested that their means were across the board lower than the control group. For example, the mean score on the Neale Analysis of Reading Ability for Reading Recovery, control, and comparison children was 36.6, 27.5, and 22.9, and the mean score of the Word Attack Skills Test was 102.5, 78.9, and 70.6. respectively.

It seems unethical not to analyze differences between the control and comparison groups based on attrition in the control group, yet analyze differences between the control and Reading Recovery group. I suspect there were significant differences at medium-term maintenance between the remaining control children and comparison children although these results were not reported. Thus, these differences would show that the control group now consisted of low-normal children who would be expected to score higher than low-performance comparison children. If this were so, then at medium-term maintenance the appropriate control group to compare to Reading Recovery would have been the comparison group rather than the control group. It would be appropriate to use the comparison group in that way because at short-term maintenance, before 15 children were removed, the control and comparison groups were shown to be no different.

Because no results were presented for the comparison children compared to either control children or Reading Recovery, I used the means and standard deviations presented in Table 9 (Center et al., 1995, p, 255) to calculate by hand t-tests comparing Reading Recovery and comparison children. Each of these results were significant including book level ($t = 5.93$, $df = 53$, $p < .001$), Burt Word Reading Test ($t = 3.91$, $p < .001$), Neale Analysis of Reading Ability ($t = 2.91$, $p < .01$), Passage Reading Test ($t = 4.50$, $p < .001$), Waddington Diagnostic Spelling Test ($t = 2.36$, $p < .005$), Phonemic Awareness Test ($t = 2.85$, $p< .05$), Word Attack Skills Test ($t = 2.85$, $p < .01$), and Woodcock Johnson comprehension ($t = 3.59$ p M<. 001). Thus, the results of medium-

term maintenance did show a continued advantage for Reading Recovery rather than the so-called washout effect. Effect sizes comparing Reading Recovery with the comparison group for six of the eight measures at medium-term were large (1.60 book level, 1.23 Neale, 1.16 passage reading, 1.12 Burt word reading, 1.04 Waddington spelling, and .91 Woodcock passage comprehension) and two were moderate (.79 for word attack and .66 for phonemic awareness). The large effect size for the standardized tests is especially noteworthy as is the moderate effect size of phonemic awareness. Thus, while the authors' conclusion that Reading Recovery is very effective at points of discontinuation and over the short term is clearly supported, their conclusion that the effects are washed out over the long run is not supported by their own data. These results instead show that Reading Recovery is effective in producing gains in both standardized reading tests and in phonemic awareness beyond what were gained by a comparison group of low-performing children. Again, a major result of this study is the demonstration of Reading Recovery effectiveness even when the researchers had a critical stance toward the intervention.

A much-discussed result of this study is the authors' inspection of individual children's performance at medium-term maintenance and consideration of whether they had indeed, a year after Reading Recovery, still remained within the normal range on a standardized test as indicated by their performance on a variety of word and passage reading, phonemic awareness, decoding, and comprehension tests. At this point 65% of Reading Recovery children, compared to 37% of the control group and 29% of the comparison group, achieved at or above the nearly- normal range on a standardized test. The authors argued from these results that approximately 30% (between the 37% and 29% found in the control and comparison groups) of Reading Recovery children would have reached normal levels of achievement without Reading Recovery. Furthermore, the authors noted that 35% of the Reading Recovery children did not achieve age- level expectations. Therefore, they argued that the current use of the Observation Survey is insufficient in identifying children who most need and will benefit from Reading Recovery. They concluded that many children are enrolled who could accelerate without the benefit of Reading Recovery; and many children are enrolled who do not accelerate. A later study addressing this same issue (Schwartz, 2005) is discussed later in this paper. It is important to note here, that Center et al. (1995) used the results of a standardized test given 1 year after Reading Recovery to determine whether children had achieved in the normal range.

The fifth major quantitative study of Reading Recovery was published in 2001 in *Scientific Study of Reading* (Chapman, Tunmer, & Prochnow, 2001). The details of this study are presented in Table 1. This study was conducted by researchers who are not associated with Reading Recovery, and one author (Tunmer) had previously conducted work critical of Reading Recovery's

approach to teaching phonological processing (Iversen & Tunmer, 1993). In the present study the authors examined children's reading and writing skills on a wide range of measures one year prior to Reading Recovery, during the Reading Recovery year, and at the middle of the year following Reading Recovery. The researchers claimed that Reading Recovery children begin school with phonological processing deficiencies, that Reading Recovery did not eliminate or reduce these deficiencies, and that performance in Reading Recovery was closely related to phonological processing skills. They concluded, "that RR children were in need of instruction in word-decoding strategies but did not receive it" (Chapman, et al., 2001, p. 173).

However, the authors' strong statements about the ineffectiveness of Reading Recovery need to be examined in light of several methodological concerns. The researchers did not use a random assignment or even quasi-experimental matching procedures. Instead, they collected data on an entire cohort of students over 2½ years, and only retrospectively identified a comparison group from this cohort against which to compare the progress of Reading Recovery students. Retrospectively the researchers determined that using scores at the end of kindergarten, they could include the lowest 20 children who never entered Reading Recovery in a control group and that this group had no significant differences on assessments with the Reading Recovery group at that point in time. Thus, they argued that both groups were equivalent. However, the students in this so-called equivalent control group must have made good progress without the intervention because if they had not, they would have entered Reading Recovery later in the year and, therefore, not appeared in either the control group or Reading Recovery group. This explains why only 20 children were included in the control group rather than an equivalent number to the Reading Recovery group. While 32 children were selected for Reading Recovery, only 20 children were identified as controls. Inspection of the data at every data point prior to Reading Recovery and after showed that the control children's mean scores were higher (and sometimes substantially so) than the Reading Recovery children although statistical analyses never showed these differences were significant.

It is also important to note that the analyses used in this study were inappropriate. The authors used a series of analyses of variance rather than multivariate analysis of covariance (therefore controlling for initial differences) despite analyzing up to 14 different measures at one data point. Further, they used Scheffes as follow-up statistics. This statistic is very conservative compared to other follow-up statistics that could have been employed. This suggests that the large differences between the Reading Recovery and control group may have been significant if a more liberal statistic had been used. It seems that the control group, which the researchers treat as a true low-performing comparison group, were instead low-normal readers. One way to determine this is to

examine whether Reading Recovery children began to close the gap between their mean performance and the mean performance of this low-normal group after receiving Reading Recovery. Recall the authors found no statistical differences between these two groups using their conservative statistic.

Table 2 presents the mean scores for the four groups included in this study on four critical variables—phonological segmentation, pseudoword reading, invented spelling, and text levels—at the end of Year 1 prior to Reading Recovery (Time 3), at the middle of the second year at the point of expected discontinuation for Reading Recovery (Time 4), and at the end of that year (Time 5). The two variables that are not likely to align with Reading Recovery

**Table 2. Means for Phonological Segmentation, Pseudoword Reading, Invented Spelling and Text Levels at Time 3 (End of Year 1), Time 4 (Middle of Year 2), and Time 5 (End of Year 2)**

| | Time 3 | Time 4 | (Gain) | Time 5 | (Gain) |
|---|---|---|---|---|---|
| Phoneme Segmentation | | | | | |
| ND[a] | 15.53 | 17.94 | (2.4) | 19.33 | (1.4) |
| PRC[b] | 8.30 | 12.79 | (4.5) | 16.41 | (3.6) |
| RR/D[c] | 5.24 | 13.13 | (7.9) | 17.48 | (4.4) |
| RR/R[d] | 1.40 | 4.40 | (3.0) | 11.40 | (7.0) |
| Pseudoword Reading | | | | | |
| ND | 54.90 | 74.21 | (19.3) | 78.26 | (4.1) |
| PRC | 20.25 | 45.74 | (25.5) | 51.41 | (5.7) |
| RR/D | 12.21 | 35.42 | (23.2) | 45.91 | (10.1) |
| RR/R | 6.40 | 16.80 | (10.4) | 20.00 | (3.2) |
| Preconventional (Invented) Spelling | | | | | |
| ND | 46.86 | 55.36 | (8.5) | 58.45 | (3.1) |
| PRC | 29.15 | 42.16 | (13.0) | 44.88 | (2.7) |
| RR/D | 16.80 | 38.21 | (21.4) | 43.96 | (5.6) |
| RR/R | 7.20 | 20.80 | (13.6) | 26.80 | (6.0) |
| Text Level | | | | | |
| ND | 13.61 | 19.35 | (5.8) | 22.57 | (3.2) |
| PRC | 6.13 | 9.92 | (3.8) | 15.00 | (5.1) |
| RR/D | 4.88 | 9.02 | (4.1) | 15.17 | (6.2) |
| RR/R | 3.75 | 5.00 | (1.2) | 9.10 | (4.1) |

Note: [a] ND normally developing, [b] PRC control group/poor reader comparison, [c] RR/D Reading Recovery children who were discontinued, [d] RR/R Reading Recovery children who were not discontinued and referred for additional support (Chapman, Tunmer, & Prochnow, 2001, Tables 3, 4, and 5, pp. 156, 160, 161).

(but are argued by the researchers to be critical components of early literacy which are not explicitly addressed in Reading Recovery) are the phoneme segmentation and pseudoword reading scores whereas the invented spelling and text level scores are associated with Reading Recovery.

As shown in Table 2, at Time 3 (end of kindergarten) the control group children (PRC) scored higher than the Reading Recovery who were discontinued (RR/D) and much higher than Reading Recovery children not discontinued, but referred on (RR/R) on every measure. At Time 4 (middle of Grade 1) PRC children still scored higher on every measure except phonemic segmentation, where the RR/D children had closed the gap. The gain score presented after Time 4 represents the gain children made during the period from the end of Year 1 (end of kindergarten) to the middle of Year 2 after the Reading Recovery intervention. In most cases RR/D made greater gains that either the normally developing or the PRC control children. Thus, the Reading Recovery children who were discontinued did make greater gains overall than either the high-normal group or the PRC (which can be considered a low-normal group).

Another issue in this study is the definition of what is considered a deficiency in phonological processing. The authors treat deficiency as scoring lower than normal-developing children on assessments of phonemic awareness and phonological processing. However, the children in the normal-developing group who were used as a benchmark for appropriate levels of phonological processing were high-average rather than typical. At the middle of Year 3, their reading ages ranged from 8.3 to 8.8 compared to their chronological age of 7.6 (Chapman, et al., 2001, p. 166). It is expected that children needing Reading Recovery would, on every test we could devise, score lower than high-normal children. However, using high-normal children to make benchmarks to determine effective levels of phonological processing is problematic. Instead, we need to specify the level of performance on phoneme segmentation and recoding activities that should be expected at the middle and end of kindergarten, and at the beginning, middle, and end-of-year first grade. Then we can determine whether Reading Recovery children begin with true deficiencies and whether, after intervention, they have reached target levels of performance and eliminated deficiencies.

The sixth quantitative study of Reading Recovery appeared in 2005 in the *Journal of Educational Psychology*. Schwartz (2005) examined both effectiveness and the efficiency issues raised in Center et al. (1995). He examined a group of children randomly selected either for first- or second-round Reading Recovery intervention to determine the percentage of children who were identified for Reading Recovery (second-round students) but who made adequate progress without Reading Reading as well as the percentage of Reading Recovery children who did not make progress despite Reading Recovery intervention. He

argued that scores at the transition from first to second round service provides a strong indication of the intervention effect for at-risk students with and without intervention services. Table 1 presents the design and results of this study.

Schwartz (2005) eliminated many design flaws of earlier studies by using random assignment of true low-performing children in the same classroom to Reading Recovery treatment during first round or to second round. Thus, this researcher acknowledged Reading Recovery's identification problem (not all children will need Reading Recovery despite qualifying for it). Using text level benchmarks at discontinuation (rather than standardized measures 1 year after Reading Recovery) he found that only 14% of children who were delayed for second-round Reading Recovery made progress toward age-level expectations compared to Center et al.'s (1995) estimate of 30%. Similarly he found only 27% of Reading Recovery children were not discontinued, indicating they needed long-term support, compared to Center et al.'s estimate of 35%. Thus, Schwartz argued that the small percentage of children who will be included in Reading Recovery but who could achieve without the one-to-one intervention was counterbalanced by the reduction in children who would need long-term services, an important role of Reading Recovery. He argued that intensive intervention is an important way to identify children with specific cognitive deficits related to the reading process compared to those who merely suffer from a lack of literacy-related experiences or appropriate instruction and who do not need long-term support services. An interesting finding is that after the first round of Reading Recovery and at the end of the year, Reading Recovery children scored as well as low- and high-average readers in phoneme segmentation.

Together the studies published in Tier 1 research journals demonstrate a converging, if controversial, picture of Reading Recovery success. More importantly for this review, the research on Reading Recovery has informed the field of literacy research in several ways, three of which are highlighted here.

1. First and foremost, this research has produced a renewed awareness of the difficulty in designing intervention studies of the lowest-performing children without major flaws that severely limit the validity of results. Figure 1 presents an overview of the design components that have emerged as a result of critiques of Reading Recovery research and their inclusion in studies published in Tier 1 research journals. Studies which include more of these design components would be considered to have higher quality. In general, Reading Recovery researchers have conducted studies with slightly higher levels of quality designs than researchers critical of Reading Recovery.

2. The second major implication from research on Reading Recovery is that consumers of research must be very careful to determine whether the conclusions drawn by the researchers are, indeed,

**Figure 1. Design Elements Found in Reading Recovery Studies Published in Tier 1 Research Journals**

| | Pinnell, 1989 | Iversen & Tunmer, 1993 | Pinnell et al., 1994 | Center et al., 1995 | Chapman et al., 2001 | Schwartz, 2005 |
|---|---|---|---|---|---|---|
| Random assignment treatment and control | X | | X | X | | X |
| One-to-one control | | X | X | | | X |
| Multivariate statistics or p controls | X | | X | X | X | X |
| Covariates to control initial levels of skill | | | X | | | |
| Control and treatment same schools | X | X | X | X | | X |
| Control and treatment same classrooms | X | X | | | | X |
| Standardized measures reading | X | | X | X | X | X |
| Measures phonemic awareness | | X | | X | X | X |
| Measures comprehension | | | X | | X | X |
| Frequency distributions | | | | | | |
| Average children at mid ranking vs. all others not Reading Recovery | | X | | | | X |
| Longitudinal beyond first grade | X | | | X | X | |
| Includes all Reading Recovery children | | | X | X | | X |
| Measures treatment fidelity | | | X | X | | |

reflected in the actual data of the study. At least two studies, studies most critical of Reading Recovery, have used dubious designs or failed to produce statistical procedures that other reviewers might have demanded.

3. Finally, research on Reading Recovery has revealed weakness in the field's definition of "deficits." In the studies currently reviewed, most researchers have treated deficits as merely scoring lower than normal-developing children. Reading Recovery has always taken the stand that the outcome of its program is a self-extending reader who can continue to achieve within the classroom. Thus, teachers work toward a specified level of achievement rather than merely making readers "better." Reading Recovery rejects the notion that getting children "better" is an appropriate goal for interventions; instead, teachers work to help children achieve a self-extending system which will allow them to take advantage of regular classroom instruction and remain within the average band of achievement.

## RESEARCH PUBLISHED IN TIER 2 RESEARCH JOURNALS

There are three kinds of studies published in Tier 2 research journals: evaluation studies of RR's effectiveness in first grade, longitudinal studies of Reading Recovery, and studies that have particularly examined the role of phonemic awareness in relation to Reading Recovery.

### Evaluation Studies of Reading Recovery: First Grade

There are numerous examples of evaluations of Reading Recovery's effectiveness in first grade including an early study conducted in New Zealand (Glynn & Cross, 1992). These researchers found variations in program implementation, early advantages for children discontinued, and a washout of effects after time. Many researchers have examined the effectiveness of Reading Recovery in the United States at the district or state level using Web-based data submission procedures produced by the National Data Evaluation Center (NDEC) at The Ohio State University, or data collected within particular states or districts, or a combination of these. For example, Ashdown and Simic (2000) examined the end-of-the-year achievement of 55,231 children including 25,601 first-grade students who received Reading Recovery services in New York during a 5-year period. Reading Recovery children's performance was compared to two controls, including 11,267 comparison children who were identified for Reading Recovery but did not receive services and 18,363 random sample children who were randomly selected children from each Reading Recovery site but were not identified as needing Reading Recovery. Children in each group

were identified as English native speakers, fluent non-native speakers (ELS), and non-native speakers with limited English proficiency (LEP). Analysis of variance demonstrated an interaction between language and sample on end of the year first-grade scores on book level. Reading Recovery children's scores at the end of the year for all three language groups were similar with little difference among English speakers, fluent ESL children, and LEP children. In contrast, the differences among the scores of these three groups of children for the comparison group were pronounced and were severe for the random sample group. The authors concluded that Reading Recovery contributed to the improvement of all children's literacy achievement but was also significant in closing the gap for English language learners. Unfortunately this study is flawed because of the lack of pre-test scores to demonstrate initial levels of performance. The comparison group, although clearly low-performing children, still were likely to be less impaired given they were not selected for Reading Recovery. Therefore, analysis of covariance would be the preferred statistical method.

Quay, Steele, Johnson, and Hortman (2001) did include a low-performing control group in their study of the effectiveness of Reading Recovery in its first year of implementation. Children in 34 schools were selected to participate in the study by randomly assigning one classroom to Reading Recovery and one classroom to control. The lowest-performing children in each classroom were assigned to the respective groups and were tested at the beginning of the year with the Observation Survey and with the Iowa Test of Basic Skills (ITBS). MANOVA analyses demonstrated no differences at pretest. All children were tested in the spring of first grade on the ITBS and the Gates-MacGinitie Reading Test. Post-test MANOVA and follow-up ANOVA's demonstrated that the Reading Recovery children outperformed the control children on four of six measures of the ITBS and all four measures in the Gates-MacGinitie. Unfortunately, children in this study were not randomly selected from the same classrooms; nonetheless, a low-performing control group was included with quasi-random assignment from the same school and standardized measures were included.

Another statewide effectiveness study examined whether Reading Recovery closed the well-publicized achievement gap between children in different racial and socio-economic groups. Rodgers and Gómez-Bellengé (2003) examined the differences in achievement of two groups of Reading Recovery children in Ohio: 5,547 children whose lessons were successfully discontinued (discontinued group), 7,234 children who received a full program (treatment group including both discontinued and not discontinued children), and 1,915 first graders who did not receive Reading Recovery (comparison group). Two subtests of the Observation Survey, Hearing and Recording Sounds in Words (HRSW) and Text Reading Level (TRL), were collected for all subjects at the

beginning and end of first grade. Raw scores and stanines were disaggregated for the three groups by race (White and African-American) and by socio-economic level (free lunch vs. regular lunch). For the normal-achieving comparison group, the stanines of the subtests for the racial groups were the same in the fall; however, in the spring African-American children scored one stanine lower than White children on both HRSW and TRL. Similarly, the normal-achieving low-income children scored two stanines below middle-income children for both subtests in the fall and remained behind by two stanines in HRSW but only one stanine in TRL in the spring.

In contrast, for the Reading Recovery group (which included both discontinued and referred children) White children began the year two stanines above African-American children in HRSW and three stanines in TRL. At the end of the year the African-American Reading Recovery children differed by only one stanine compared to White children on both subtests. The results were similar for low-income versus middle-income children although differences were wider at the beginning and end; nonetheless, the gap decreased. For discontinued children the difference between African-American and White children in the fall were two stanines (HRSW) or three stanines (TRL). However, in the spring African-American children completely closed the gap and scored in the same stanine as White children for HRSW, while they were only one stanine behind in TRL (compared to a three-stanine difference in the fall). The results were very similar for children in different income levels.

This study showed that for typical first-grade children, Whites and African-Americans begin the year very close in performance; however, by the end of the year they begin to experience a gap in achievement. In contrast, children receiving Reading Recovery (including those who were not discontinued) began the year with large differences in literacy achievement favoring White and middle-income children. However, at the end of the year these differences were significantly decreased. Unfortunately, this study only included data from the Observation Survey.

## Longitudinal Evaluation Studies

Other studies published in Tier 2 research journals have examined the longitudinal effects of Reading Recovery including the effects on children who are English language learners receiving Reading Recovery in English or in Spanish through Descubriendo la Lectura (DLL), the reconstruction of Reading Recovery for children whose initial reading instruction is in Spanish. For example, Rowe (1995) in a large-scale study involving over 5,000 children, examined student, teacher, and school factors influencing reading achievement in Australia. This researcher found that the mean score of Reading Recovery children was lower than that of a comparison group of children in second through the sixth grade. However, Rowe noted that the lower edge of the range

of scores (10th percentile) for Reading Recovery after first and second grade was actually higher for Reading Recovery children than for all the other children. In addition, Reading Recovery's highest edge of range (90th percentile) was the same for three out of eight cohorts of children that were included in the study. These results clearly demonstrate the importance of moving beyond merely reporting mean scores to examining the range of scores for all children. Rowe's results demonstrated that Reading Recovery children were no longer merely clustered at the very bottom of the range but had expanded into all percentile ranks. This trend persisted through sixth grade. Other studies of Reading Recovery found similar results.

For example, Askew and Frasier (1994) compared 54 discontinued Reading Recovery and 53 randomly selected second graders' text reading levels, fluency, and retellings. MANOVA analyses revealed that the only difference among the children was in the fluency measure of pacing in which Reading Recovery read slightly slower than other children. The three measures of retelling and other two measures of fluency were not significantly different. Unfortunately, text level means were not subjected to statistical procedures; rather the researchers concluded that the mean of Reading Recovery children fell within the average band of the randomly selected group of children (defined as .5 standard deviation above and below the mean). This study did not include standardized measures, the full range of children included in Reading Recovery, or a low-performing control group; nonetheless, it examined children's comprehension which has been cited as a weakness in research on the effects of Reading Recovery (e.g., Hiebert, 1994).

Later Askew et al., (2002) examined children's performance on several measures of reading including text level, Gates-MacGinitie scores in vocabulary and comprehension, and scores on the Texas Assessment of Academic Skills (TAAS) through the fourth grade. These researchers calculated the percentage of children whose text levels were at or above grade-level expectations (95%), at or above the 4th stanine on the Gates-MacGinitie (64%), or were at or above the passing rate on the state reading assessment (85%). The percentage of discontinued Reading Recovery children's performance at or above grade level expectations was similar to all other children (98% at or above text level, 84% at stanine 4 or above on the Gates-MacGinitie, and 90% passing the state reading assessment). A small sample of children who were not discontinued but completed full programs was also examined. Surprisingly, the percentage of these children who could read at or above grade level increased from 17% in second grade to 50% in fourth grade.

Ruhe and Moore (2005) examined the reading and writing scores of 1,250 Reading Recovery children on the fourth-grade MEA, the statewide standardized test taken by all children in Maine, compared to 14,286 other fourth-grade children in the state. Reading Recovery children included those who were

discontinued from Reading Recovery, were not discontinued but were referred for additional services, and did not complete the program (the school year ended prior to their completing the program, they moved, or they were otherwise removed from the program). The control group consisted of all children not included in Reading Recovery.

These researchers found that discontinued students' average reading and writing scores fell within the average achievement band, defined as one-half standard deviation above and below the average state score, whereas referred and incomplete children's average scores did not. A majority of discontinued children's scores (57%) fell within or above the state average band. Similar results were found for the writing test. Only 10% of discontinued Reading Recovery children failed to meet any grade-level expectations in reading, and 11% failed to do so in writing.

Escamilla, Loera, Ruiz, & Rodríguez (1998) found similar results with children learning to read in Spanish. These researchers demonstrated that DLL children could read at higher Spanish text levels than a randomly selected group at both second and third grades. On the SABE-2 Spanish Reading Achieve Test, DLL students performed the same as their randomly selected peers. However, a majority of children (75% to 94%) scored at or above the average band of performance in second and third grade on these two tests. Neal and Kelly (1999) found similar results, although this study suffers from several design flaws including lack of use of multivariate analyses.

## Phonemic Awareness and Reading Recovery

A few studies have examined Reading Recovery's effect on increasing children's phonemic awareness or whether entry-level skill in phonemic awareness affects children's success in discontinuing from Reading Recovery. Iversen and Tunmer (1993) found that all children entering Reading Recovery or modified Reading Recovery had very low levels of phonemic awareness and all children benefited from instruction and outperformed a control group. Stahl, Stahl, and McKenna (1999) reported similar effects. Reading Recovery children and a low-performing control group entered with a low level of phonemic skills. Reading Recovery students made better gains in 16 weeks than control children, although the difference was small. Sylva and Hurry (1996) reported that Reading Recovery children made better progress on the Neale Analysis of Reading Ability than children in a phonological instruction intervention, although this study failed to provide results of statistical analyses.

Two studies have shown that, in general, children who are discontinued tend to have higher initial levels of phonemic awareness than children who are not discontinued (Center, et al., 1995; Chapman, et al., 2001). These authors suggest that phonemic awareness ought to both be taught within Reading Recovery more explicitly and used as a factor in selecting children to be

included in the program. However, the results of a study by Spector and Moore (2004) on the use of phonemic awareness at entry as a predictor of success in Reading Recovery suggests a different interpretation. In this study 135 Reading Recovery children taught by 37 teachers were given the Observation Survey and a test of phonological processing (Yopp-Singer Test of Phoneme Segmentation) along with a test of verbal short-term memory and rapid automatized naming at entry into and exit from Reading Recovery. The researchers examined differences in entry scores for children who were eventually discontinued compared to children who were not. Further, they investigated the effects of segmentation, verbal memory, and rapid naming on the odds for being discontinued.

Overall the students demonstrated similar skills on the Observation Survey. However, there were significant differences between the two groups on segmentation and verbal memory. All children entered with very low levels of phonemic awareness and lower than normal levels of verbal memory. Many children could not segment consistently the first phoneme in words, considered the typical performance of end-of-the-year kindergarten children. The main finding was that although the mean performance of discontinued children was higher than not-discontinued children, there were considerable numbers in both groups who could not perform any segmentation. Logistic regression using phonemic awareness and verbal memory correctly classified, based on entire level scores, 69% of the students who did not discontinue and 65% of the children who discontinued. Of the 33% who were not correctly classified, 16% of the children who were expected to discontinue due to higher levels of phonemic awareness and verbal memory failed to do so. Similarly, 17% of the children, who had lower levels of phonemic awareness and verbal memory and who were not expected to discontinue, did so.

This study is noteworthy for two reasons. First, the researchers identified typical levels of segmentation that would be expected in normal-developing children and compared Reading Recovery children's performance to these benchmarks. Thus, they moved beyond using depressed mean scores as a definition of phonemic awareness deficits. Second, they clearly showed that while phonemic awareness levels are in general related to success in Reading Recovery, a significant portion of children manage to be successful despite very low levels of phonemic awareness, whereas a significant portion of children with much higher entry levels do not capitalize on this narrow band of skills to accelerate reading. These researchers call for research to discover how some Reading Recovery teachers are able to take children predicted to fail and allow them to be successful. Overall, Reading Recovery is successful teaching phonemic awareness as demonstrated in Center et al. (1995) and Iverson and Tunmer (1993); yet it appears some Reading Recovery teachers are more skillful at this instruction than others. Unfortunately, the researchers did not report results of phonemic awareness assessments at the end of the program, so it is

not possible to determine the range of growth in this measure after Reading Recovery for both discontinued and not discontinued children.

### Summary of Research Published in Tier 2 Research Journals

Studies published in Tier 2 research journals suffer from several design flaws that are inherent in retrospective longitudinal studies including lack of random assignment of children to Reading Recovery and a true lowest-performing comparison group. Other concerns include failing to use more powerful multivariate statistics including covariates and failing to include reports of all Reading Recovery children's performance along with discontinued children. Figure 2 on the following page presents an overview of the design components included in the studies published in Tier 2 research journals. Nonetheless, these studies have strengths not found in studies published in Tier 1 research journals. Many of the studies from Tier 2 research journals present data from very large samples of children and report the range of children's performance as well as mean performance.

For example, Schmitt and Gregory (2005) conducted a longitudinal examination of Reading Recovery children and children who had not been selected for Reading Recovery. The children were randomly selected from a large data pool and all children were selected from the same classrooms. At the beginning of first grade all children were tested on the Observation Survey and the Gates-MacGinitie. The Gates-MacGinitie and a book level measure were administered to both groups of children in second, third, and fourth grades. These researchers examined the percentage of each cohort that read at or above grade level as well as the frequency distribution of all children across text levels at each grade level. Inspection of only the mean scores in this study suggests that Reading Recovery *as a group* may have scored lower than a random sample of all other children. However, inspection of the frequency data across various text levels suggests that high percentages of Reading Recovery children scored above grade-level expectations. Only a few Reading Recovery children (and inspection of the data suggests that about the same number of other children) scored below the grade-level expectation. The real difference between Reading Recovery and other children was that higher numbers of Reading Recovery children scored at text levels just above the grade-level expectation, whereas higher numbers of other children scored at the very highest text levels. Thus, the mean scores would be different, but only considering means would mask the meaningful and important result that Reading Recovery children had moved beyond the lowest 20% of the scores into a range of reading levels clustered just above grade-level expectations. Unfortunately, only discontinued children were examined in this study.

These reports, using large samples of children including control groups, provide further converging evidence of the effectiveness of Reading Recovery.

**Figure 2. Design Elements Found in Reading Recovery Studies Published in Tier 2 Research Journals**

| | Askew & Frazier, 1994 | Rowe, 1995 | Neal & Kelly, 1999 | Brown et al., 1999 | Ashdown & Simic, 2000 | Quay et al., 2001 | Askew et al., 2002 | Escamilla et al., 1998 | Rodgers & Gómez-Bellenge, 2003 | Ruhe & Moore, 2005 | Schmitt & Gregory, 2005 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Low-performing control group | | | | | X | X | | | | | |
| Random assignment to treatment and control | | | | | | X | | | | | |
| Multivariate | X | X | | | | X | | | | X | |
| Analysis covariance | | | | | | | | | | | |
| Same schools | X | | | | X | X | X | X | X | | X |
| Same classrooms | | | | | | | | | | | X |
| Standardized measure | | X | | X | | X | X | X | | X | X |
| Phonemic awareness measure | | | | | | | | | | | |
| Comprehension measure | X | X | | | | X | X | X | | | X |
| Avg. group with mid-range scores vs. all others | | | | | | | | | | | |
| 80% all others control | X | X | X | | X | | X | X | X | X | X |
| True longitudinal (beyond first grade) | 2nd | 6th | | 5th | | | 4th | 3rd | | 4th | 4th |
| Includes all RR (at least full 20 weeks); not just discontinued | | | X | X | X | X | X | | X | X | |
| Treatment fidelity | | | | | | | | | | | |
| Percentage met expectation/benchmarks or above reported | | | | X | | | X | X | | X | X |
| Frequency distribution of all scores reported | | X | | X | | | X | | | | X |
| Includes ELL | | | X | | X | | | X | | | |

Nonetheless, there are weaknesses to these studies, many of which had been discussed in studies published in Tier 1 research journals. First, in most studies the control group is not specified; rather the control group is defined as the top 80% of the school population rather than a purposefully selected sample of children scoring near the average band. A second issue emerging from these studies is that researchers do not sufficiently discuss the significant differences between Reading Recovery children and others even when these differences are reported. For example, Rodgers and Gómez-Bellengé (2003) found that closing the gap for all children on TRL was more difficult than for HRSW; an interesting finding that suggests that teaching children phonemic awareness is far easier than getting it applied in real reading. The result most striking to me, again not discussed, is the difference in achievement in Reading Recovery children *at the beginning* of the year. It is interesting that African-American and low-income children in Reading Recovery started the year behind their White and middle-class cohorts who are also in Reading Recovery; however, only 853 African-American children were served in Reading Recovery in the study year, compared to 4,453 White children. A discussion of the possible reasons for this difference would have strengthened the research.

In the next section I describe studies that have synthesized the research on Reading Recovery. Although I discuss these syntheses separately, their impact needs to be considered in relationship to the research reports previously reviewed.

## SYNTHESES OF RESEARCH PUBLISHED IN TIER 1 RESEARCH JOURNALS

Three early critiques of Reading Recovery (Nicholson, 1989; Robinson, 1989; Center & Wheldall, 1992) focused primarily on design flaws in Clay's research and on other early unpublished evaluations of Reading Recovery outside the United States. In 1993, the first research synthesis of Reading Recovery appeared in a Tier 1 research journal. Wasik and Slavin (1993) compared Reading Recovery to four other one-to-one tutorial programs (although two of the five programs were intended for second grade as well as first-grade children). These researchers calculated overall effect sizes based on all available data (including data presented by Pinnell in 1989 and data that would later be published in Pinnell et al. in 1994). They demonstrated that Reading Recovery had a range of effect sizes for TRL from moderately high to high (.72 and .78, p. 185; 1.50 and .75, p. 186) in the first year of implementation with diminishing results to small effect sizes 2 years after implementation (.14 and .25, p. 185). The effect sizes for Success for All were calculated on the mean of several measures of reading including word attack, oral reading, and silent reading. During the first year of implementation mean effect sizes ranged from

high to moderate (1.01, 2.37, .84, 1.83 at one site and .55, .87, and .55 at a second site, p. 190.) Effect sizes for Prevention of Learning Disabilities in the first year of implementation ranged from high to small (.85 and .33 for oral reading and .16 for total reading, p. 192) while effect sizes for Wallach Tutoring Program ranged from large to moderate (.50, .64, .60, 1.8, on word recognition and .75 on total reading, p 193) and were moderate for 30 minutes of Programmed Tutorial Reading (. 57 on vocabulary and .53 on comprehension, p. 194). It would not be appropriate to compare Reading Recovery's outcomes to the other programs due to differences in the ages of children, comprehensiveness of approaches, and overall program goals. Nonetheless, this study demonstrated that one-to-one tutorial programs, including Reading Recovery, do produce significant effects.

Wasik and Slavin (1993) raised three important points in their review. First, they argued that research on the effects of one tutoring model or another are suspect to bias in measurement; researchers select assessments matched to the instruction they provide, biasing their students to be more successful than students not receiving such instruction. In addition, they argued that most researchers take a pragmatic approach to demonstrating the success of their program rather than attempt to craft research which would provide theoretical insights into why the program is effective. Finally, they raised the question of whether research has sufficiently addressed the question of fidelity to program. Most researchers have not spent sufficient time observing actual instruction to be confident that teachers are providing instruction that reflects that actual goals and instructional procedures defined in the program. Wasik and Slavin point out that Pinnell et al.'s (1994) research is exemplary in this regard. Similarly, they argued that quality of teaching aside from mere program fidelity would also affect outcomes and has not been addressed sufficiently in research. Again, Wasik and Slavin noted that while Pinnell et al. did not focus on documenting the effects of a range of teaching quality, they did remark on its existence within programs.

Hiebert (1994) took a different approach to synthesizing the research on Reading Recovery. She examined whether Reading Recovery was sufficient in making appreciable changes within an age cohort in a single school. She considered an age cohort to be 72 children who would be in three first-grade classrooms typically served by one FTE Reading Recovery teacher (who are actually two .5 FTE teachers). Hiebert used data drawn from reports obtained from three Reading Recovery University Training Centers: The Ohio State University, Texas Woman's University, and University of Illinois at Champaign. Data were collected from 1984–85 until 1991–92, where the information was available. She argued, based on her inspection of the data, that teachers on average discontinued 11 children per year. Then she determined the percentage of Reading Recovery children who continued to perform on grade level in

fourth grade as obtained from one unpublished study (DeFord, Pinnell, Lyons & Place, 1990) which was available at that time. In the DeFord et al. study, 36% of the Reading Recovery children were reading text on grade level and 41% performed at average or above levels on the Woodcock Reading Mastery Test. Thus, Hiebert argued that in fourth grade only 4 children (40% of Reading Recovery children on grade level times 11 discontinued Reading Recovery children) out of 72 would actually have benefited from Reading Recovery. Hiebert argued that "data from the three primary RR sites and from the longitudinal study (DeFord et al., 1990) produce an unconvincing scenario of the effects of RR on an age cohort" (1984, p. 23). She further argued that even if the percentage of other Reading Recovery cohorts that maintain grade-level achievement increases, the number of children in each age cohort will still remain low because of the number of children who can be served in the program.

In fact, more recent research has shown that considerably higher percentages of discontinued Reading Recovery children do maintain grade-level expectations through the third and fourth grade. For example, Askew et al. (2002) found that 95% of 116 Reading Recovery children in fourth grade were reading grade-level texts, 63% were within the average band on a standardized test, and 90% passed the Texas reading assessment. Ruhe and Moore (1995) found that 57% of 726 discontinued children in fourth grade were reading within or above the average band of the Maine reading assessment as established by 14,285 children. Escamilla et al. (1998) found that 75% of 274 Spanish speaking third graders were reading on or above grade level and 79.2% scored at or above grade level on a Spanish reading achievement test. Schmitt and Gregory (2005) found that 90% of 277 discontinued children in fourth grade were reading at or above grade level and 80% scored at or above an average band on a standardized test. Thus, even 3 years after receiving Reading Recovery services, a much larger portion of children are still operating within grade-level expectations than Hiebert estimated.

Despite Hiebert's criticisms of research on Reading Recovery (lack of measures of comprehension and using comparison groups with only a mediocre program with low levels of professional development), she highlights important components of effective beginning reading instruction found in Reading Recovery: time spent reading and writing, attention to phonemic awareness, deliberate instruction, high expectations, and experimenting with letter-sound relationships in writing.

Shanahan and Barr (1995) also synthesized a variety of research related to Reading Recovery. First, they examined differences in Reading Recovery as it is implemented in the United States compared to New Zealand and noted three major differences. Children in New Zealand have substantially more academic instruction prior to first grade than is the case in the U.S., and thus children in

New Zealand enter Reading Recovery with higher levels of knowledge. Children are discontinued at a faster rate in New Zealand (not surprising given their higher initial entry scores), and children in New Zealand are in regular classrooms with instruction more aligned with Reading Recovery than is the case in the U.S.

Second, Shanahan and Barr (1995) summarized the results of five research studies available to date on Reading Recovery (including studies already reviewed by Wasik and Slavin). They argued that these results must be approached with caution because they omit data from a large percentage of children who entered Reading Recovery but did not complete the programs. That is, they reported that in 1991, one state failed to include data on 22% of the total sample of Reading Recovery children including 9% who ended the year without complete programs, 7% who were placed in special education and removed from the program, and 6% who moved from the school. Shanahan and Barr argued that Reading Recovery's effectiveness must be judged not just on the children who discontinue or who fail to discontinue given an entire program, but on all children enrolled in Reading Recovery, a procedure now routinely followed in reporting Reading Recovery data. Further, they argued that the superior performance of the Reading Recovery children could be an artifact of regression toward the mean.

Third, the authors summarize the effect sizes from Wasik and Slavin's (1993) study of tutorial programs in reading. They also calculated the effect sizes from a small-scale study of a restructured program for low-performing Title I children (Hiebert, Colt, Catto, & Gury, 1992) and regular classroom children (Taylor, Short, Frye, & Schearer, 1992). The effect sizes for these two studies ranged from 1.16 to .48. They compared these to effect sizes for Reading Recovery (.72 and .78 as reported in Wasik and Slavin, 1993, p. 185).

In conclusion, Shanahan and Barr (1995) argued that Reading Recovery is effective for low-performing first graders, although other small-group instructional strategies may also be effective (based on two small-scale studies which included children performing beyond the bottom 20%). They concluded that school systems would not be wise to place all their dollars in one-to-one tutoring, but would need to continue to support special education and services to children beyond the first grade in order to maintain accelerated learning. In fact, Reading Recovery has been used as an effective device for reducing the total number of children in special education (e.g., Lyons & Beaver, 1995) and insuring that those children who are referred are indeed special needs children rather than children merely lacking exposure to highly effective instruction (Schwartz, 2005).

While these three reviews included effect size comparisons, they were not able to use meta-analysis techniques due to the lack of sufficient number of studies with pre- and post-test data on treatment and control children.

However, two more recent reviews have used these methods to provide a research synthesis. The first synthesis (Elbaum et al., 2000) examined the effectiveness of one-to-one tutoring programs for at-risk readers, including Reading Recovery. This study included 29 studies that were "published or available between 1975 and 1998" (p. 606) including studies previously reviewed in earlier studies. One effect size was calculated for each sample yielding 41 effect sizes. Effect sizes were aggregated by variables such as who delivered the instruction (teacher, volunteer, college student), type of measure (e.g., comprehension, spelling, writing, text level), and type of intervention (Reading Recovery, other interventions using teachers, interventions delivered in first grade). Of interest here is that Reading Recovery had an overall effect size of .66 which was significantly higher than the mean effect size of .29 for matched interventions in first grade.

Elbaum et al. (2000) also compared effect sizes for two additional studies (Acalin, 1995; Evans, 1996) that the authors claimed compared Reading Recovery instruction to other intervention instruction delivered in small groups. Acalin's study, an unpublished master's thesis, involved kindergarten through fourth-grade students. Some students were taught Reading Recovery and others a small- group intervention based on the Orton-Gillingham approach. The effect size in this study was -.12. Evans' (1996) study, an unpublished doctoral dissertation, took place in one first-grade classroom in which eight of the lowest-performing children were identified in kindergarten and four children were randomly selected to receive Reading Recovery in first grade. The other four children were taught by the classroom teacher (Evans) in a modified Reading Recovery approach comprised of shared and guided reading and rereading of familiar books along with writing. Although no statistical analyses were conducted in this case study, the children selected for small-group instruction scored higher than the children selected for Reading Recovery at pre-test on Letter Identification (LI), Concepts About Print (CAP), Writing Vocabulary (WV), and HRSW. The effect size for this study was .05.

To further examine differences between Reading Recovery and small-group instruction, Elbaum et al. (2000) calculated the effect sizes of the one-to-one instruction in Reading Recovery presented in the Pinnell et al. (1994) study compared to the small-group instruction condition. They combined all three of the one-to-one conditions (RR, RS, and DISP) compared to the RWG condition and found an effect size of -.12. Using the two unpublished studies and their own calculation of effect sizes on the Pinnell et al. study, the researchers concluded, "together the findings from the available small-group comparison indicate that when highly qualified teachers implement a well-designed intervention the academic benefit to students is the same, whether students are taught individually or in a group of 2 to 6 students" (Elbaum et al., 2000, p. 616).

Clearly, these researchers did not consider fidelity of program before claiming researchers were investigating Reading Recovery. No program delivered kindergarten through fourth grade could be Reading Recovery. Thus, Acalin's (1995) study was not a true comparison of Reading Recovery and a small-group intervention. Evans' (1996) results could be due to teacher effect—only one Reading Recovery and classroom teacher participated in the study. Finally, a more direct measure of the effectiveness of one-to-one Reading Recovery versus small group in Pinnell et al. (1994) would be to compare effect sizes of the RR with RWG. Based on data in Table 7 (Pinnell et al., 1994, p. 27), the effect size for these two groups on TRL was .92; for HRSW the effect size was.32; and for Gates- MacGinitie the effect size was .12. Thus, the effect sizes Elbaum et al. reported as reflecting differences between Reading Recovery and small-group instruction are flawed.

Elbaum et al.'s (2000) study suffers from additional flaws which are better revealed after considering the latest meta-analysis on Reading Recovery in the U.S. (D'Agostino & Murphy, 2004). D'Agostino and Murphy "conducted a comprehensive search of the literature utilizing ERIC, PychInfo, and dissertation Abstracts databases" (p. 27) in order to locate studies in which treatment fidelity was reported (children actually received Reading Recovery instruction which was not considered in Elbaum et al.'s study), and where sufficient data was reported to compute effect sizes. They found 36 studies published on or before 1998 that met all their criteria for inclusion. However, only 11 studies met the additional criteria of having both pre- and post-test data for both treatment and control groups. Thus, these researchers conducted two levels of meta-analyses: the first level was for all studies and the second level was for only research which met higher levels of design quality. In the second level analyses the researchers statistically controlled for pre-test status of all children. They disaggregated the data for discontinued children, not discontinued children, and all Reading Recovery children combined.

In the first analysis, the authors calculated effect sizes comparing Reading Recovery children with low-performing controls and regular children for standardized measures and for each of six measures on the Observation Survey. The effect sizes on standardized tests administered in first grade compared to low-performing controls was .48 for discontinued children and .32 for all children in Reading Recovery. At second grade, the effect size compared to low-performing children on standardized tests jumped to .66 for discontinued children and .63 for children in Reading Recovery overall. The results on the Observation Survey ranged from high effect sizes for TRL (2.78) WV (1.40), and HRSW (1.12) for discontinued children compared to low-performing controls. Even when compared to regular children, effect sizes for discontinued children remained moderate (.51 for TRL, .44 for WV, and .34 for HRSW).

Not discontinued children's effect sizes were usually negative compared to all children, discontinued children, and normal children except on WV (.33) and HRSW (.41) compared to other low-performing children. Effect sizes at pre-tests were also calculated and showed that Reading Recovery children began the studies with lower abilities than the low-performing controls.

In the second analysis the authors used higher-quality studies and calculated weighted effect sizes taking into account pre-test differences. These analyses demonstrated lower effect sizes for standardized tests (.27 for discontinued children and .20 for all children in Reading Recovery). However, large effect sizes, although lower than found in the first analysis, were still found for TRL (1.54 for discontinued and 1.66 for all children), WV (.95 for discontinued children and 1.01 for all children) and HRSW (1.04 for discontinued children and .94 for all children). Thus, this second analysis, where Reading Recovery and control children had similar pre-test scores, and where pre-test status was controlled, demonstrated strong effects for Reading Recovery on several Observation Survey measures and small effects on standardized measures. Due to the results of the second analysis in which pre-test scores were not different and initial status was controlled, the authors argued that the results of Reading Recovery can not be merely caused by regression toward the mean. Further, they included in their analyses all Reading Recovery children, not just discontinued children, and found positive effects even for children who were not discontinued. Thus, the results of previous studies cannot be attributed to omission of data in which only successful Reading Recovery children are included, or to regression to the mean.

It is interesting to compare the two meta-analyses on the studies used in the analyses and in the conclusions drawn from the research. D'Agostino and Murphy (2004) located 36 studies of Reading Recovery available on or before 1998; Elbaum et al. (2000) located 10 studies during the same time period. Only five studies are included in both reviews. D'Agostino and Murphy located 6 technical reports, 6 published articles, 22 ERIC documents, 1 dissertation, and 1 paper presented at a conference. Elbaum et al. found 3 published papers, 3 ERIC documents, 2 dissertations, and 2 papers presented at conferences. It is interesting that Elbaum et al. used 3 ERIC documents but failed to locate the other 19 documents that D'Agostino and Murphy used in their analyses.

Elbaum et al. (2000) concluded in their summary that several Reading Recovery studies selectively drop students from the program and that this represents a "particularly pernicious form of participant attrition in which the researchers selectively remove participants from a study based precisely on the participants' failure to respond adequately to the treatment (p. 616). Further, they claimed that "overall the findings of this meta-analysis do not provide support for the superiority of Reading Recovery over other one-to-one reading

interventions" (p. 617). They argue that the positive effects of Reading Recovery can be attributed to omitting students and using biased measures (Clay's Text Reading Level). Yet, this conclusion is not supported by their data. They found that Reading Recovery had an effect size of .66 compared to .29 for all other matched one-to-one interventions. In contrast, D'Agostino and Murphy (2004) concluded that, "to date, the bulk of available evidence indicated that RR has had positive effects on participating students across outcomes designed for the program and external to it, and that results of more rigorously designed studies seemed to converge with this conclusion" (p. 35–36). Further, they reported data for all Reading Recovery children so their effect sizes cannot be attributed to participant attrition as claimed by Elbaum et al.

## ISSUES RAISED BY RESEARCH ON READING RECOVERY

Together the quantitative research on Reading Recovery raises several issues of concern to literacy researchers.

### Difficulties in Investigating Lowest-Performing Children

Investigators who aim to examine the effectiveness, efficiency, and sufficiency of instruction for the very lowest-performing children face both ethical and design decisions which I have previously discussed. If the very lowest-performing children are selected for intervention because of fundamental beliefs about the need to serve those most in need, then children selected for the control group are not likely to be as low performing. When control group children are not as low-performing as children in the intervention group, then positive results for interventions can be explained by regression to the mean and negative results can be explained by arguing that the control group is really a low-normal group rather than a true low-performing comparison.

As I have shown in this review, both of these competing explanations have been applied to many of the studies of Reading Recovery published in Tier 1 and 2 research journals. One way to mitigate the lack of random assignment of children to intervention and control groups because of ethical reasons, is to randomly assign schools to intervention and control status depending on availability of services. However, Borman and Hewes (2003) argued that schools should not be forced to adopt innovations, including intervention programs. Rather, they should willingly support the intervention. One way to handle such situations is to assign schools to early and later implementation. This is the model Schwartz (2005) used in his study where two of the lowest-performing children were randomly selected to either first- or second-round Reading Recovery entry. Future research on early intervention programs would be wise to adopt such models.

## Readers Beware

Studies published in Tier 1 research journals as well as some studies published in Tier 2 research journals include design and methodological flaws including inappropriate selection of comparable control groups, inappropriate statistical analyses, omitted analyses, overstated results, and use of a small number of studies or unpublished reports to draw key conclusions. For example, Center et al. (1995) did not compare a control group with a comparison group due to attrition in the control group, but compared that same control group to Reading Recovery. The results of this flawed analysis that failed to show a Reading Recovery effect have been quoted in several other studies also critical of Reading Recovery (e.g., Shanahan & Barr, 1995; Elbaum et al., 2000). Elbaum et al. used two unpublished studies to conclude that one-to-one instruction in Reading Recovery is not more effective than small-group instruction. They even inappropriately combined non-Reading Recovery treatments with the Reading Recovery group in Pinnell et al.'s (1994) data to calculate effect sizes used to argue that Reading Recovery was not better than the group instruction treatment. When only considering the difference between the appropriate Reading Recovery group and the small-group treatment, I found the effect sizes were moderate and positive.

Stanovich and Stanovich (2003) argued that researchers, and especially teachers, need to be informed consumers of research. Informed readers consider whether a large body of well- designed studies on the same phenomena consistently supports one or more competing theories. Competing theories to Reading Recovery include the notion that small-group instruction is equally as effective as one-to-one instruction and interventions must include explicit phonemic awareness and phonological recoding. Competing interpretations of research include conclusions that Reading Recovery is only effective for a small percentage of children and has only immediate effects which later wash out so that intermediate-age children served by Reading Recovery are no longer different from their low-performing peers and do not perform within the class average.

Taking into account that no one research study could ever address all competing theories or competing conclusions drawn from research or solve all design issues, informed readers of research would ponder the evidence of a large body of research to determine whether research begins to converge and consistently support one or more competing theories. In the case of Reading Recovery, five studies out of six included in my review of studies published in Tier 1 research journals have demonstrated the superiority of Reading Recovery in general. (Chapman et al., 2001, did not report supportive results at any point in time.) This is true even for studies where researchers are critical of Reading Recovery. Thus, there is converging evidence that Reading Recovery is effective.

There is less convergence for the other competing theories or interpretations. For example, one study refuted the competing theory that small-group instruction is as effective as Reading Recovery (Pinnell et al., 1994) and one (flawed) research synthesis claimed to support small-group instruction (Elbaum et al., 2000). Two studies presented differing results regarding Reading Recovery's ability to serve large percentages of children effectively (Center et al., 1995; Schwartz, 2005). Two studies have demonstrated that Reading Recovery is effective in teaching phonemic awareness and segmentation (Center et al., 1995; Iversen & Tunmer, 1993). One study published in Tier 2 research journals demonstrated that Reading Recovery produces gains in phonemic awareness beyond other interventions (Stahl et al., 1996). Many studies published in Tier 2 research journals support the interpretation that large percentages of Reading Recovery children, primarily children discontinued, sustain their gains and remain within the average band of achievement. Thus, there is much converging evidence that Reading Recovery is effective and children successfully discontinued do remain within the average band of achievement. There is less evidence supporting competing theories or interpretations of research findings.

## Even Small Effect Size Matters, But Only in Combination With Other Evidence

In an era when instructional practices are expected to be based on scientifically based research, professionals are expected to consider research evidence in order to make appropriate decisions regarding instruction, curriculum, and assessment. Slavin (2005) argued that "a strategy of using the findings of rigorous research as a basis for policy and practice depends on the existence of a substantial body of research that identifies practical, replicable models for school and classroom reform" (p. 7). He reviewed several areas of educational practices that have been studied using high-quality designs including randomized longitudinal studies of Reading Recovery and DLL (Escamilla, 1994) using effect sizes as indications of the quality of effectiveness. The effect sizes he reported for Reading Recovery were those previously reported by Wasik and Slavin (1993) and for DLL ranged from .97 to 1.61. In a more comprehensive analysis, D'Agostino and Murphy (2004) reported effect sizes of .48 and .32 on standardized tests at the end of Year 1 for discontinued and all Reading Recovery children compared to other low-performing children. At the end of Year 2 effect sizes increased to .66 for discontinued children and .63 for all children. Using only the highest-quality research, as recommended by Slavin (2005), they found effect sizes of .27 and .20 for discontinued and all children on standardized tests at the end of Year 1.

Thus, considering only the effects of Reading Recovery on standardized tests and only using the most rigorous research, Reading Recovery's effect sizes

ranged from .20 to .27. Cohen (1988) specified that an effect size of .20 or smaller should be considered a small effect which would be expected in fields such as clinical psychology. Lipsey and Wilson (1993) argued that educational treatments with small effects of even .10 should not be considered trivial. To put such values in perspective, Borman, Hewes, Overman, and Brown (2003) evaluated 49 studies of Direct Instruction, 42 studies of Success for All, and 10 studies of School Development Program, among 232 studies of 29 Comprehensive School Reform Models. They reported a 95% confidence level for Direct Instruction of .17 to .25 with a mean effect size of .21. The range for School Development Program was .10 to 20 with a mean effect size of .15. The range reported for Success for All was .16 to .21 with a mean of .18. The range for all 29 Comprehensive School Reform Models was .09 to .15 with a mean effect size of .12. Reading Recovery's mean effect size of .20 to .27 compares favorably to effect sizes of far more comprehensive models of school reform ranging from means of .15 to .21 for rigorously studied models and .09 to .15 for all models.

It is important to keep in mind that the small effect sizes for Reading Recovery on standardized measures are by far the most conservative of its effects. On more directly related measures on the Observation Survey, the effect sizes are very large. But more importantly and beyond mere effect sizes, Reading Recovery has never been intended to merely get children "better" than another method. Instead, the program has been intended to change the learning trajectory of children so they move out of the lowest-performing group and enter and remain in a range of achievement levels closer to and approximating the normal band of achievement. Studies published in Tier 2 research journals suggest that a range of between 57% (Ruhe & Moore, 2005) and 80% (Schmitt & Gregory, 2005) of discontinued children do so on high-stakes state assessments or other standardized tests at the fourth grade.

## Consider the Source of Critique

The reason I choose to review the quantitative studies of Reading Recovery was because early in my reading I discovered that this was contested territory. Several researchers have expressed concerns with the instructional procedures used in Reading Recovery or the measures used to identify children for Reading Recovery intervention (Iversen & Tunmer, 1993; Hiebert, 1994; Center et al., 1995; Elbaum et al., 2000; Chapman et al., 2001). While it is not possible to identify the actual sources of critique of Reading Recovery, there seem to be three underlying concerns: the sources of funds used to support Reading Recovery, the control over who should best instruct lowest-performing readers, and the most effective programs used to identify children who need special education. Perhaps easiest to understand is Hiebert's (1994) motivation to critique Reading Recovery and its siphoning off of Title I funds for a small

group of children given that she had earlier published a study of the effects of an intervention program delivered in what was then called Chapter I (Hiebert, et al., 1992). Thus, it is not surprising that she would be concerned about Title I funding. However, meta-analyses of Title I have shown it to have very little effect (Borman & D'Agostino, 2001). Similarly, it is easy to understand educational psychologists' positions on the central and isolated role of phonemic awareness and phonological recoding given they have played prominent roles in this research.

More concerning are the implications of critique of Reading Recovery from some researchers who work within the domain of special education: for example, Vaughn (in Elbaum et al., 2000), and Center et al., 1995. One explanation for their critique is suggested in the emerging new method of identifying children who qualify for special education which is explored in four articles published in a recent issue of *Reading Research Quarterly* (2006, p. 92–128). These articles were devoted to current issues in special education and reading instruction and shed light on new methods of defining who will address the needs of low-performing readers in the early grades and how their needs will be met. As discussed in these articles, the Individuals with Disabilities Education Improvement Act (IDEIA) which became effective in July 2005, included a new method to identify children with learning disabilities called "response to intervention" (RTI).

RTI is an alternative method of identifying children for special education. First, a group of at-risk children are identified using performance on assessments. Using established criteria, students who perform below expectations are selected for more intense classroom intervention or small-group instruction. When students are not responsive to this intense instruction (again based on assessment), then they can be assessed for inclusion in special education. School systems can use up to 15% of their special education budget for early intervention activities as defined by RTI procedures, and teachers providing the instruction do not necessarily need to be special educators. It is assumed this method will reduce the number of children in special education by eliminating children whose only difficulty is lack of exposure to effective instruction.

It is interesting that RTI procedures match those of Reading Recovery. In Reading Recovery the lowest-performing children are selected for intervention; after intervention, children are returned to the regular classroom or are referred on for special services, often special education. Thus, it seems clear that research which could be used to bolster the effectiveness of using RTI to identify special needs children would include research on Reading Recovery.

It is telling that three of the four articles on RTI do not refer to Reading Recovery nor its research studies. The fourth article (McEneaney, Lose, & Schwartz, 2006), did make the connection to Reading Recovery, calling it contingent teaching, and made explicit Reading Recovery's mission to reduce

the number of children eventually relegated to special education and to strengthen the certainty that only children with cognitive deficits would be identified for services in this program. This article provided a welcome sign of collaboration among classroom teachers, intervention teachers including Reading Recovery teachers, and special educators.

On the one hand it is possible that Reading Recovery teachers, literacy educators, and special educators will work more closely together to support the literacy learning of at-risk children including special needs children; RTI provides the opportunity to do this. On the other hand, I find it problematic that the authors of the three other papers did not find connections between models of responsiveness to instruction suggested by researchers who work with the domain of special education (e.g., Vaughn, Linan-Thompson, & Hickman, 2003) and Reading Recovery. On the positive side, a possible sign of cooperation is implied in these articles. For example, Fuchs and Fuchs (2006) applaud the "larger role for reading specialists (in the Responsive Teaching Intervention), which in turn might affect pre- and inservice professional development activities conducted by universities and school districts." They argued that this new provision (RTI) "has implications for the number and type of children identified, the kinds of educational services provided, and who delivers them" (Fuchs & Fuchs, 2006, p. 93).

## CONCLUSION

The research on Reading Recovery provides insights for all literacy professionals, especially in light of recent innovations in special education. Gersten and Dimino (2006) stated that "the purpose of RTI is not only to provide early intervention for students who are at risk for school failure but also to develop more valid procedures for identifying students with reading disabilities" (p. 100). This statement is similar to arguments that Reading Recovery advocates would make. It is important that the converging positive results of the large body of research on Reading Recovery do not become marginalized as critics refer again and again to the few studies (with clear flaws) that do not show positive effects or do not attend to the research at all. If early literacy professionals allow the positive effects found in research on Reading Recovery to be minimized, then so too will be the role of literacy researchers in defining the nature of literacy curriculum, instruction, and assessment. It is critical that literacy professionals remain in leadership roles, as well as collaborate with others including special educators, in identifying at- risk readers and shaping instruction that will best fit their needs. RTI and the new provisions of IDEIA will allow Reading Recovery teachers, along with all early literacy leaders, a new role in collaborating to help our lowest-performing children reach their potential.

**Author's Note**

The author thanks Dr. Robert Schwartz and two anonymous reviewers for their thoughtful comments, and Dr. Richard G. Lomax for his statistical advice.

## REFERENCES

Acalin, T. A. (1995). *A comparison of Reading Recovery to Project READ*. Unpublished doctoral dissertation, California State University, Fullerton.

Ashdown, J., & Simic, O. (2000). Is early literacy intervention effective for English language learners? Evidence from Reading Recovery. *Literacy Teaching and Learning: An International Journal of Early Reading and Writing, 5*, 27–42.

Askew, B. J., & Frasier, D. F. (1994). Sustained effects of Reading Recovery intervention on the cognitive behaviors of second grade children and the perceptions of their teachers. *Literacy, Teaching, and Learning: An International Journal of Early Literacy, 1*(1), 87–107.

Askew, B. J., Kaye, E., Frasier, D. F., Mobasher, M., Anderson, N., & Rodríguez, Y. G. (2002). Making a case for prevention. *Literacy Teaching and Learning: An International Journal of Early Reading and Writing, 6*(2), 43–73.

Borman, G. D., & D'Agostino, J.V. (2001). Title I and student achievement: A quantitative synthesis. In G.D. Borman, S.C. Stringfield, & R.E. Slavin (Eds.), *Title I: Compensatory education at the crossroads* (25–57). Mahwah, NJ: Erlbaum.

Borman, G. & Hewes, G. (2003). Long-term effects and cost effectiveness of Success for All. *Educational Evaluation and Policy Analysis, 24*(2), 243–266.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73*(2), 125–230.

Briggs, C., & Young, B. K. (2003). Does Reading Recovery work in Kansas? A retrospective longitudinal study of sustained effects. *The Journal of Reading Recovery, 3*(1),59–64.

Brown, W., Denton, E., Kelly, P., & Neal, J. (1999). Reading Recovery effectiveness: A five-year success story in San Luis Coastal Unified School District. *ERS Spectrum: Journal of School Research and Information, 17*(1), 3–12.

Center, Y., & Wheldall, K. (1992). Evaluating the effectiveness of Reading Recovery: A critique. *Educational Psychology, 12*(3/4), 263–275.

Center, Y., Wheldall, K., Freeman, L., Outhred, L., & McNaught, M. (1995). An evaluation of Reading Recovery. *Reading Research Quarterly, 30*(2), 240–263.

Chall, J. S. (1989). Learning to read: The great debate 20 years later—a response to "Debunking the great phonics myth." *Phi Delta Kappan, 70*, 521–538.

Chapman, J. W., Tunmer, W. E., & Prochnow, J. E. (2001). Does success in the Reading Recovery program depend on developing proficiency in phonological-processing skills? A longitudinal study in a whole language instructional context. *Scientific Studies of Reading, 5*(2), 141–176.

Clay, M. M. (1966). *Emergent reading behavior*. Unpublished doctoral dissertation, University of Auckland, New Zealand.

Clay, M. M. (1987). Implementing Reading Recovery: Systemic adaptations to an educational innovation. *New Zealand Journal of Educational Studies, 22*, 35–58.

Clay, M. M. (1991). *Becoming literate: The construction of inner control*. Auckland, New Zealand: Heinemann.

Clay, M. M. (1993). *Reading Recovery: A guidebook for teachers in training*. Portsmouth, NH: Heinemann.

Clay, M. M. (2006). *An observation survey of early literacy achievement* (revised 2nd ed.). Portsmouth, NH: Heinemann.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cohen, S. G., McDonnell, G., & Osborn, B. (1989). Self-perceptions of at risk and high achieving readers: Beyond Reading Recovery achievement data. In S. McCormick & J. Zutell (Eds.). *Cognitive and social perspectives for literacy research and instruction: Thirty-Eighth Yearbook of the National Reading Conference* (pp. 117–122). Chicago, IL: National Reading Conference.

Cox, B. E., Fang, Z., & Schmitt, M. C. (1998). At-risk children's metacognitive growth during the Reading Recovery experience: A Vygotskian interpretation. *Literacy Teaching and Learning: An International Journal of Early Reading and Writing, 3*, 55–76.

D'Agostino, J. V., & Murphy, J. A. (2004). A meta-analysis of Reading Recovery in United States schools. *Educational Evaluation and Policy Analysis, 26*(1), 23–38.

DeFord, D. E., Pinnell, G. S., Lyons, C., & Place, A. W. (1990). *The Reading Recovery follow-up study, 11*. Columbus: The Ohio State University.

Elbaum, B., Vaughn, S., Hughes, M. T., & Moody, S. W. (2000). How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology, 92*(4), 605–619.

Escamilla, K. (1994). Descubriendo La Lectura: An early intervention literacy program in Spanish. *Literacy, Teaching and Learning: An International Journal of Early Literacy 1*(1), 57–70.

Escamilla, K., Loera, M., Ruiz, O., & Rodríguez, Y. (1998). An examination of sustaining effects in Descubriendo la Lectura programs. *Literacy Teaching and Learning: An International Journal of Early Reading and Writing, 3*(2), 59–81.

Evans, T. L. P. (1996). *I can read deze books: A qualitative comparison of the Reading Recovery program and a small-group reading intervention*. Unpublished doctoral dissertation, Auburn University, Alabama.

Forbes, S., & Szymczuk, M. (2003). *Iowa's study of sustained effects of the Reading Recovery intervention* (Technical Report). Iowa City: University of Iowa, Reading Recovery Center of Iowa.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to Response to Intervention: What, why, and how valid is it? *Reading Research Quarterly, 41*(1), 93–99.

Gersten, R., & Dimino, J.A. (2006). RTI (response to intervention): Rethinking special education for students with reading difficulties (yet again). *Reading Research Quarterly, 41*(1), 99–108.

Glynn, T., & Cross, T. (1992). Reading Recovery in context: Implementation and outcome. *Educational Psychology, 12*(3/4), 249–262.

Glynn, T., Crooks, T., Bethune, N., Ballard, K., & Smith, J. (1989). *Reading Recovery in context*. Wellington, NZ: Department of Education.

Gómez-Bellengé, F. X. & Thompson, J. R. (2005). *U.S. norms for tasks of an observation survey of early literacy achievement*. (Rep. No. NDEC 2005–02). Columbus: The Ohio State University, National Data Evaluation Center. http://www.ndec.us

Goodman, K. (1986). *What's whole in whole language*? Portsmouth, NH: Heinmann

Heibert, E. H., Colt, J., Catto, S., & Gury, E. (1992). Reading and writing of first-grade students in a restructured Chapter 1 program. *American Educational Research Journal, 29*, 545–572.

Hiebert, E. H. (1994). Reading Recovery in the United States: What difference does it make to an age cohort? *Educational Researcher, 23*, 15–25.

Iversen, S., & W. E. Tunmer. (1993). Phonological processing skills and the Reading Recovery program. *Journal of Educational Psychology, 85*(1), 112–126.

Jaggar, A. M., & Simic, O. (1996). *A four-year follow-up study of Reading Recovery children in New York state: Preliminary report* (Technical Report). New York: New York University, School of Education.

Lipsey, M. W., & Wilson, D. B. (1993). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Lyons, C. A., & Beaver, J. (1995). Reducing retention and learning disability placement through Reading Recovery: An educationally sound, cost-effective choice. In R. L. Allington & S. A. Walmsley (Eds.), *No quick fix: Rethinking literacy instruction in America's elementary schools* (pp. 61–77). Newark, DE: International Reading Association.

McEneaney, J. E., Lose, M. K., & Schwartz, R. M. (2006). A transactional perspective on reading difficulties and Response to Intervention. *Reading Research Quarterly, 41*(1), 117–128.

Neal, J. C., & Kelly, P. R. (1999). The success of Reading Recovery for English language learners and Descubriendo la Lectura for bilingual students in California. *Literacy Teaching and Learning: An International Journal of Reading and Writing, 4*(2), 81–108.

Nicholson, T. (1989). A comment on Reading Recovery. *New Zealand Journal of Educational Studies, 24*(1), 95–97.

O'Connor, E., & Simic, O. (2002). Effect of Reading Recovery on special education referrals and placements, *Psychology in the Schools, 39*(6), 635–646.

Pinnell, G. S. (1989). Reading Recovery: Helping at-risk children learn to read. *The Elementary School Journal, 90*(2), 161–183.

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*(1), 9–39.

Quay, L. C., Steele, D. C., Johnson, C. I., & Hortman, W. (2001). Children's achievement and personal and social development in a first-year Reading Recovery program with teachers in-training. *Literacy Teaching and Learning: An International Journal of Early Reading and Writing, 5*, 7–25.

Rasinski, T. V. (1995). Commentary on the effects of Reading Recovery: A response to Pinnell, Lyons, DeFord, Bryk, and Seltzer. *Reading Research Quarterly, 30*(2), 264–270.

Robinson, V. (1989). Some limitations of systemic adaptation: The implementation of Reading Recovery. *New Zealand Journal of Educational Studies, 24*(1), 35-45.

Rodgers, E. M., & Gómez-Bellengé, F. X. (2003). Closing the achievement gap in Ohio with Reading Recovery. *The Journal of Reading Recovery, 3*(1), 65–74.

Romei, G. (2002). *An assessment of longitudinal outcomes of the Reading Recovery intervention in Maine with at-risk populations*. Unpublished doctoral dissertation, University of Maine, Orono.

Rowe, K. J. (1995). Factors affecting students' progress in reading: Key findings from a longitudinal study. *Literacy, Teaching and Learning: An International Journal of Early Literacy, 1*, 57–76.

Ruhe, V., & Moore, P. (2005, Winter). The impact of Reading Recovery on later achievement in reading and writing. *ERS Spectrum, 23*(1)20–30.

Rumbaugh, W., & Brown, C. (2000). The impact of Reading Recovery participation on students' self-concepts. *Reading Psychology, 21*, 13–30.

Schmitt, M. C., Askew, B. J, Fountas, I. C., Lyons, C. A., & Pinnell, G. S. (2005). *Changing futures: The influence of Reading Recovery in the United States*. Worthington, OH: Reading Recovery Council of North America.

Schmitt, M. C., & Gregory, A. E. (2005). The impact of an early literacy intervention: Where are the children now? *Literacy Teaching and Learning: An International Journal of Early Reading and Writing, 10*(1), 1–20.

Schwartz, R. M. (2005). Literacy learning of at-risk first-grade students in the Reading Recovery early intervention. *Journal of Educational Psychology, 97*(2), 257–267.

Shanahan, T. (1987). Book review: The early detection of reading difficulties, by Marie M. Clay, *Journal of Reading Behavior, 19*, 117–119.

Shanahan, T., & Barr, R. (1995). Reading Recovery: An independent evaluation of the effects of an early instructional intervention for at-risk learners. *Reading Research Quarterly, 30*(4), 958–996.

Slavin, R. (2005). *Evidence-based reform: Advancing the education of students at risk*. Washington, DC: Center for American Progress.

Spector, J. E., & Moore, P. (2004). Does phonological processing distinguish between students who are more or less responsive to Reading Recovery? *Literacy Teaching and Learning: An International Journal of Early Reading and Writing, 8*(2), 1–25.

Stahl, K. A. D., Stahl, S., & McKenna, M. C. (1999). The development of phonological awareness and orthographic processing in Reading Recovery. *Literacy Teaching and Learning: An International Journal of Early Reading and Writing, 4*(1), 27–42.

Stanovich, P. J. & Stanovich, K. E. (2003). *Using research and reason in education: How teachers can use scientifically based research to make curricular & instructional decisions*. Washington, DC: U.S. Department of Education.

Sylva, K., & Hurry, J. (1995). Early intervention in children with reading difficulties: An evaluation of Reading Recovery and a phonological training. *Literacy, Teaching and Learning: An International Journal of Early Literacy, 2*(2), 49–68.

Taylor, B. M., Short, R. A., Frye, B. J., & Shearer, B. A. (1992). Classroom teachers prevent reading failure among low-achieving first-grade students. *The Reading Teacher, 45*, 592–597.

Thompson, G. B., Fletcher-Finn, C. M., & Cottrell, D. S. (1991). *Sources of grapheme-phoneme correspondence knowledge during the acquisition of reading and spelling*. Unpublished manuscript.

Vaughn, S., Linan-Thompson, S., & Hickman, P. (2003). Response to instruction as a means of identifying students with reading/learning disabilities. *Exceptional Children, 69*, 391–409.

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*(2), 178–200.

Yopp, H. K. (1988). The validity and reliability of phonemic awareness tests. *Reading Research Quarterly, 23*, 159–177.

Reading Recovery® Council
of North America