Efficacy of a First-Grade Responsiveness-to-Intervention Prevention Model for Struggling Readers

Jennifer K. Gilbert  
Donald L. Compton  
Douglas Fuchs  
Lynn S. Fuchs  
*Vanderbilt University, Nashville, TN, USA*

Bobette Bouton  
*The University of Georgia, Athens, USA*

Laura A. Barquero  
Eunsoo Cho  
*Vanderbilt University, Nashville, TN, USA*

**ABSTRACT**

This randomized control trial examined the efficacy of a multitiered supplemental tutoring program within a first-grade responsiveness-to-intervention prevention model. Struggling first-grade readers \((n = 649)\) were screened and progress monitored at the start of the school year. Those identified as unresponsive to general education Tier 1 \((n = 212)\) were randomly assigned to receive Tier 2 small-group supplemental tutoring \((n = 134)\) or to continue in Tier 1 \((n = 78)\). Progress-monitoring data were used to identify nonresponders to Tier 2 \((n = 45)\), who were then randomly assigned to more Tier 2 tutoring \((n = 21)\) or one-on-one Tier 3 tutoring \((n = 24)\). Tutoring in Tier 3 was the same as in Tier 2 except for the delivery format and frequency of instruction. Results from a latent change analysis indicated nonresponders to Tier 1 who received supplemental tutoring made significantly higher word reading gains compared with controls who received reading instruction only in Tier 1 (effect size = 0.19). However, no differences were detected between nonresponders to Tier 2 who were assigned to Tier 3 versus more Tier 2. This suggests more frequent 1:1 delivery of a Tier 2 standard tutoring program may be insufficient for intensifying intervention at Tier 3. Although supplemental tutoring was effective in bolstering reading performance of Tier 1 nonresponders, only 40% of all Tier 2 students and 53% of Tier 2 responders were reading in the normal range by grade 3. Results challenge the preventive intent of short-term, standard protocol, multitiered supplemental tutoring models.

The responsiveness-to-intervention (RTI) approach to preventing academic difficulties has gained momentum in research and school communities, representing one of the more prominent education reforms in recent years. While originally grounded in special education law (Individuals with Disabilities Education Act, 2004) as a legal alternative to the IQ-discrepancy approach for identifying students with learning disabilities, RTI has evolved into a general education prevention system aimed at improving performance of students at risk for poor academic outcomes (see Kratochwill, Volpiansky, Clements, & Ball, 2007; Lembke, McMaster, & Stecker, 2010). Current RTI models consist of multitiered systems mirroring those developed in mental health (O’Connell, Boat, & Warner, 2009) and medicine (Mrazek & Haggerty, 1994). These prevention systems are predicated on the assumption that early intervention prior to the onset of significant problems can place an individual on a developmental trajectory associated with positive long-term outcomes (e.g., Snow, Burns, & Griffin, 1998). In addition, there is a general belief that preventive programs have greater efficacy and are more cost effective than remedial programs that wait for problems to emerge before treatment is initiated (e.g., O’Connell et al., 2009; Torgesen, 2002).
In the area of reading, school-based RTI models have adopted a preventive framework, operationalized as a multilevel system with three distinct and increasingly intensive tiers (Bradley, Danielson, & Hallahan, 2002; Donovan & Cross, 2002): primary, secondary, and tertiary. The role of primary prevention is to reduce the number of new cases of an identified condition or problem in the population, such as ensuring that all students are exposed to high-quality instruction in the general education classroom. In RTI Tier 1 (a form of primary prevention), all students are screened for initial risk using universal screening, participate in generally effective reading instruction in the general education classroom, and are monitored for rate of individual reading growth. Those whose level of performance and/or rate of improvement are dramatically below those of their peers are designated as at risk for poor reading outcomes and move to a second tier.

Secondary prevention is concerned with reducing the number of existing cases (i.e., prevalence) of an identified condition or problem in the population by promoting skill acquisition known to promote typical skill development. In RTI Tier 2 (a form of secondary prevention), the extra effort is focused on students at high risk of developing difficulties but before any serious long-term deficit has emerged. Students in this tier receive preventive small-group tutoring, and their progress is again monitored. Those who are responsive to Tier 2 continue in it and eventually return to Tier 1 only; at that point, they are deemed disability free. Students unresponsive to Tier 2 are assumed to have an intrinsic deficit that prevents them from benefiting from generally effective reading instruction (Vaughn & Fuchs, 2003). Failure to respond to Tier 2 instruction signals a need for Tier 3.

Tertiary prevention is concerned with reducing the complications associated with identified problems or conditions. In RTI Tier 3 (a form of tertiary prevention), instruction typically involves more intensive individualized instruction. Failure to respond adequately to Tier 3 prevention signals possible disability and the need for special education evaluation so well-trained school personnel can provide instruction according to an individualized education program. Programs, strategies, and interventions at Tier 3 have an explicit remedial or rehabilitative focus because students who enter this tier show behaviors related to those of students with disabilities. If students demonstrate inadequate progress under tertiary prevention, they may need treatment in the form of supplemental instruction that is specially designed (e.g., special education). The duration of typical RTI programs within the research literature last anywhere from 8 to 24 weeks (Vaughn, Denton, & Fletcher, 2010) and tend to be shorter than those used in mental health.

The effectiveness of early prevention systems such as RTI hinges on the system’s ability to accurately identify students most at risk for future difficulty (e.g., Compton et al., 2010; Compton, Fuchs, Fuchs, & Bryant, 2006) and to provide timely delivery of increasingly intensive academic intervention based on the needs of the student (L.S. Fuchs & Fuchs, 1998).

Research on RTI Models

Prior to RTI reform efforts, research showed that preventive tutoring and early remediation (delivered by adults) can improve the long-term outcomes of young students who are at risk for or diagnosed with reading difficulty (e.g., Blachman et al., 2004; Foorman et al., 1997; Mathes, Howard, Allen, & Fuchs, 1998; Torgesen et al., 1999; Vellutino et al., 1996). Such studies are important building blocks for the RTI framework because they provide evidence that early intervention can substantially improve the skills of poor readers. However, studies that incorporate multiple tiers of instruction with progress-monitoring–based decision making are still needed to assess the efficacy and preventive intent of increasingly intensive multiltiered instruction that is distinctive of RTI models.

Because we targeted first-grade students in general education classrooms for early supplemental intervention in the present study, the following review of studies is limited to those involving students who (a) were identified as at risk for reading difficulty, (b) were enrolled in first-grade general education classrooms, and (c) received tutoring that was supplemental to classroom instruction. Because two theoretical viewpoints exist as to whether the supplemental instruction approach should be scripted and standardized (standard protocol) versus individualized (problem solving), we believe it is important to state up front that six of the seven studies we identified used a standard protocol, scripted tutoring program (i.e., Case et al., 2010; Kerins, Trotter, & Schoenbrodt, 2010; Mathes et al., 2005; McMaster, Fuchs, Fuchs, & Compton, 2005; Vadasz, Jenkins, & Pool, 2000; Wanzek & Vaughn, 2008), and one used an individualized approach (i.e., O’Connor, Harty, & Fulmer, 2005).

There are trade-offs between the two approaches (D. Fuchs & Fuchs, 2006): flexibility of instructional timing and content, individualization of instruction, and reliance on expertise of the individual making decisions about assessment and instruction (high for problem solving, low for standard protocol) as opposed to the ability to attribute student outcomes to particular types of instruction, efficiency, and reliance on published research (high for standard protocol, low for problem solving). There has been a paucity of research directly
comparing these approaches. Thus, one approach cannot be deemed superior to the other. In this study, we chose the standard protocol approach primarily to exercise experimental control so outcomes would be generalizable to the content, timing, and method of the supplemental instruction program presented here rather than to tutors’ judgment and their use of a host of assessments and potential instructional programs.

Research evaluating multtier prevention systems on the reading outcomes of first-grade students has produced equivocal results. We first present null effects and then those deemed statistically significant. Case et al. (2010) randomly assigned students who had poor fall-of-first-grade word identification fluency (WIF) scores to receive 30 minutes of supplemental tutoring three days per week from a graduate research assistant or continue usual classroom instruction (two hours of phonics, guided reading, and spelling). The supplemental tutoring consisted of lessons in letter–sound correspondence, decoding practice, spelling, sight-word recognition, vocabulary, fluency, and comprehension. At the end of 11 weeks of small-group tutoring, the authors reported significantly better growth on decoding fluency (of words taught during intervention) and spelling for tutored students. No significant differences were found on norm-referenced tests of word reading, however. Results suggest that Tier 2 tutoring is capable of fostering faster growth for struggling readers, but it does not provide evidence that faster growth translates into higher scores on measures less aligned with treatment.

Kerins et al. (2010) found results similar to those of Case et al. (2010). Kerins et al. randomly assigned poorly performing first-grade readers to receive supplementary small-group tutoring provided by the speech-language pathologist and special educator or remain in their classrooms without tutoring. Tutoring consisted of phonological awareness instruction and multisensory phonics instruction, both standardized and scripted. Supplemental tutoring lasted 17 weeks, totaling 16.5 hours. Although significant growth on running records and on phoneme blending, identification, and segmentation was detected for tutored students, a pretest–posttest analysis revealed no group differences on any of these measures or norm-referenced standardized tests of word reading.

McMaster et al. (2005) also reported null results for supplemental tutoring. The authors randomly assigned first-grade students deemed unresponsive to Tier 1 classwide Peer Assisted Learning Strategies (PALS; provided by classroom teachers) to 13 weeks (three day a week for 35 minutes) of one of three standardized, scripted supplemental treatments: continued PALS, modified PALS, or individual tutoring by trained graduate research assistants. The same materials were used in all three supplemental treatments, and all three were implemented for the same amount of time; the primary differences among them were pace, proportion of time spent in each activity, and adult versus peer tutor in the third condition. Instruction components included phonological awareness, letter–sound correspondence, decoding, text fluency, and sight-word recognition. Although treatment effect sizes were moderate, no statistically significant between-group differences were detected on posttest word reading measures.

Among research efforts producing stronger treatment effects, Vadasy et al. (2000) provided struggling first-grade students with Tier 2 standard protocol, scripted supplemental tutoring four days per week for the entire school year. Tutoring focused on phonological skills, letter–sound correspondence, decoding, writing, spelling, and text fluency and was provided by trained research assistants (mostly parents). Vadasy et al. reported significant treatment effects on norm-referenced tests of word reading compared with students randomly assigned to control.

Wanzek and Vaughn (2008) also reported significant treatment effects on standardized measures when experimentally manipulating time in Tier 2. Students who were unresponsive to Tiers 1 and 2 in the fall of grade 1 were randomly assigned during the spring semester to 13 weeks of 30 minutes of Tier 2 tutoring per day, 60 minutes of Tier 2 tutoring per day, or no Tier 2 tutoring. Tutoring was provided by graduate research assistants and research associates. The standard protocol, scripted lessons revolved around building skill in phonemic awareness, phonics, sight-word recognition, fluency, vocabulary, and comprehension. Although the 60-minute group scored significantly higher on word attack than did the no-tutoring group, no differences were detected on measures of word reading. In addition, the authors reported no differences on any measure between 30 minutes of Tier 2 and none, or between 30 minutes and 60 minutes of Tier 2. They posited that in addition to amount of instruction, type of instruction and instructional grouping should be taken into account to intensify tutoring.

In examining the effect of type of Tier 2 tutoring, Mathes et al. (2005) reported significant treatment effects across two theoretically different Tier 2 instructional approaches, proactive reading (PR) or responsive reading (RR), but not between them on selective reading outcomes. After two batteries of screening tests, students at risk for poor reading outcomes were randomly assigned to one of three conditions where they received instruction daily for eight months: enhanced classroom (EC) instruction only, EC + PR, or EC + RR. The two supplemental programs were implemented by the classroom teachers for 40 minutes per day. Both programs provided explicit instruction in letter–sound correspondence, decoding,
et al. (2005) reported that 99% of students in PR and
result of supplementary tutoring. At posttest, Mathes
portions of students who achieved adequate reading as a
criterion in our study.

SD reading because it represents performance within 0.5
above the 30th percentile fall in the normal range of
sal definition, Torgesen suggests that students reading
138
100% of Tier 2 students and 69.50% of Tier 1 students
as scoring above the 25th percentile. For decoding,
range by the end of the year. Normal range was defined
100% of students who were unresponsive to Tier 1 instruction,
students who were in the normal range.

Finally, McMaster et al. (2005) found that of the
students who were unresponsive to Tier 1 instruction,
and therefore at risk for future reading difficulties, 55%
were reading at or above the 30th percentile on untimed
word identification and/or decoding at posttest. Across
these three studies, we see that supplementary or multi-
tiered tutoring is capable of preventing reading diffi-
culty for some but not all students in the same year that
they receive tutoring. To adequately assess prevention,
though, longer term follow-up is needed (Torgesen, 2000).

Purpose of the Study
Overall, previous studies suggest that first-grade multi-
tier prevention has the potential to improve reading out-
comes for students who are at risk for long-term reading
difficulties and at least somewhat robust to instruction-
al approach but may not have the power to close the
achievement gap with classroom peers. However, previ-
ous studies of first-grade multitier prevention programs
have not exercised experimental control (i.e., random
assignment at Tiers 2 and 3) in exploring the relative
benefits of Tier 3 versus Tier 2 for nonresponders to
Tiers 1 and 2, nor have they assessed the long-term
preventive intent of RTI-based intervention programs.
Additionally, a need exists to assess whether response to
instruction achieves the preventive intent of RTI mod-
els. Therefore, we seek to extend the literature by
employing random assignment at both decision points
within a three-tier RTI model and collecting follow-up
data through third grade.

In this study, we ask three research questions:

1. Is 14 weeks of standard protocol supplemental tu-
tering (Tier 2 and/or Tier 3) effective for students
identified as unresponsive to Tier 1?
generally high ratings of fidelity to instructional assistants have been shown to implement tutoring with literacy coaches, or special educators, trained research graduate research assistants. Although not a substitute approaches in D. Fuchs, Mock, Morgan, & Young, 2003). Although we acknowledge the debate surrounding use of this pedagogical framework versus an individualized, problem-solving approach (see descriptions of both approaches in D. Fuchs, Mock, Morgan, & Young, 2003). Tutoring in our study was conducted by trained graduate research assistants. Although not a substitute for well-trained, highly experienced classroom teachers, literacy coaches, or special educators, trained research assistants have been shown to implement tutoring with generally high ratings of fidelity to instructional protocols (90%, Case et al., 2010; 97%, McMaster et al., 2005; 89%, Vadasy et al., 2000) and ratings of teaching quality (2.47 out of 3, Wanzek & Vaughn, 2008). We note that among the studies discussed in the previous section on RTI, two recruited school personnel (speech-language pathologist and special educator, Kerins et al., 2010; classroom teachers, Mathes et al., 2005) to implement supplemental tutoring to nonresponders to Tier 1; yet, effects were mixed. Our tutors were provided 20 hours of training, which is well within the range of what other RTI studies have provided (25 hours, Case et al., 2010; one day, McMaster et al., 2005; 14 hours, Vadasy et al., 2000; 15 hours, Wanzek & Vaughn, 2008).

Although the prevailing thought in the RTI literature is that students who have not responded to Tiers 1 and 2 need more intensive instruction, there is no research to specify what instruction is most effective for these particular students. D. Fuchs and Fuchs (2006) suggested five ways to increase intensity: more systematic and scripted instruction, higher frequency, longer duration, smaller student groupings, or instructors with more expertise. We chose three of the five: systematic and scripted instruction, higher frequency, and smaller student groupings. These three components were implemented in a study by Denton, Fletcher, Anthony, and Francis (2006) as they attempted to remediate a group of third through fifth graders who did not show adequate response in Mathes et al.’s (2005) study.

Tier 3 instruction consisted of an intensive 16-week program; instruction in decoding (using a packaged program) was implemented for two hours a day in the first eight weeks, and instruction in fluency (using a packaged program) was implemented for one hour a day in the second eight weeks. The skills that were addressed between Tiers 2 and 3 appear to be similar, but the dose of intervention was greater in Tier 3 (60–120 minutes/day vs. 40 minutes/day), and the student to teacher ratio was smaller (2:1 vs. 3:1). The authors reported that 12 of the 27 students showed a significantly positive response to Tier 3. How these results relate to Tier 3 for younger students is unclear, but we take the moderately positive effects associated with increasing dose and decreasing teacher to student ratio while using a standard program to mean that perhaps increasing instructional intensity in this way might prove beneficial for first graders in need of Tier 3 instruction.

Finally, the length of our program was relatively short at 14 weeks of instruction; however, this too is in the range of what other studies have provided (11 weeks, Case et al., 2010; 17 weeks, Kerins et al., 2010; 13 weeks, McMaster et al., 2005; 8 weeks to 3 years, O’Connor et al., 2005; whole school year, Vadasy et al., 2000; 13 weeks, Wanzek & Vaughn, 2008).

To accurately model variance in pretest to posttest gains across treatment and control groups, we employed
latent change models within a structural equation model framework (see McArdle, 2009) using four indicators of word reading ability at each time point and also accounted for dependency in the data due to clustering. Latent change scores alleviate many of the problems associated with traditional gain scores. Word-level reading was chosen as the outcome of interest because improving word reading was the focus of our tutoring instruction, and it is an important building block for literacy development (Torgesen, 2000).

Method

Participants and Procedures

First-grade students were recruited for participation in two consecutive years (i.e., two cohorts). The same schedules of testing, treatment, and inclusion criteria were used both years. These students were in 11 schools (six of which served student populations from which more than 95% of the student body was economically disadvantaged) and 69 classrooms across the two cohorts. The district in which these schools were located mandated kindergarten instruction for all students; reading was part of the kindergarten curriculum. The schema for student selection and treatment implementation is shown in Figure 1. In the fall, teachers nominated the lower half of readers in their classrooms. Of these 1,035 students, 628 provided positive signed parental consents and verbally assented to participate.

Screening

We screened these 628 students with three 1-minute measures: two WIF lists and a rapid letter-naming screen (RLN) assessment. With WIF, students are presented with a single page of 100 high-frequency words randomly sampled from high-frequency words on the Dolch preprimer, primer, and first-grade level lists (L.S. Fuchs, Fuchs, & Compton, 2004). Students have one minute to read as many words as possible. If they hesitate on an item for four seconds, the examiner prompts them to proceed. Test–retest reliability exceeds .90. With RLN, the speed at which students name an array of the 26 letters is measured. The score is the number of letters correctly identified in one minute. Test–retest reliability of RLN exceeds .95. Scores were prorated if a student named all items in less than one minute. Performance on the three screening measures was as follows: RLN, $M = 42.51$ letters/minute, $SD = 13.03$; WIF1, $M = 20.05$ words/minute, $SD = 15.63$; WIF2, $M = 17.18$ words/minute, $SD = 17.05$.

Selecting Potentially At-Risk Students

To identify an initial pool of students potentially at elevated risk for poor reading outcomes if they went without intervention (i.e., remained in Tier 1), we applied latent class analysis (Nylund, Asparouhov, & Muthén, 2007) to the three screening measures. The purpose of such an analysis is to obtain model-based latent (unobserved) categories of students who are performing
similarly on the three screening measures. Models were developed and evaluated using Mplus version 6 (L.K. Muthén & Muthén, 1998–2010). We tested models that yielded two, three, or four classes (also known as categories) to see which best fit the data structure in each cohort because models with different numbers of categories are not nested (i.e., a one-category model cannot be considered a subset of the two-category model), and chi-square change tests are not usable. As such, B.O. Muthén and Muthén (2000) recommend four criteria for selecting the optimal number of latent categories (for details of each, see Connell & Frye, 2006): Bayesian information criteria, Lo–Mendell–Rubin likelihood ratio test (Lo, Mendell, & Rubin, 2001), entropy, and usefulness and interpretability of latent categories.

Results presented in Table 1 indicated that across both cohorts the Bayesian information criteria and LMR LRT values support a three-category model over two- or four-category models. (The entropy estimates were adequate for all models.) In addition, a three-category model provides a useful and interpretable set of categories in which to identify an at-risk group. Performance in the categories across the cohorts was similar on the screening measures. It was clear that students in category 1 had substantially lower RLN and WIF performance compared with students in categories 2 and 3. Thus, students in category 1 from both cohorts (cohort 1: n = 223; cohort 2: n = 214) were designated as potentially at risk for poor reading outcomes (for a total sample, n = 437) and selected for progress monitoring (PM). Students in categories 2 and 3 were excluded from further analyses.

**Identifying Tier 1 Nonresponders and Pretesting**

To minimize the number of students who were falsely identified as unresponsive to Tier 1, we conducted six weeks of PM with the 437 students identified as potentially at-risk. At each assessment wave, two alternate forms of WIF were administered, and the scores were averaged. Each student’s WIF performance over the six weeks fit to a line using an ordinary least-squares regression procedure. Slope was expressed as the estimated number of words per minute gained per week and level as the estimated number of words per minute read at week 6. Of the 437 students selected for PM, 10 moved during PM, and one was removed from the study because of scheduling conflicts.

McMaster et al. (2005) found the dual-discrepancy approach (considering deficits in both slope and intercept) to be a reliable and valid indicator of nonresponse. Therefore, students were rank ordered based on intercepts and slopes, and approximately the bottom half of students (n = 232, 53.09%) were identified as unresponsive to Tier 1 (i.e., classroom instruction) and thus classified as at risk for reading difficulty. Responders performed better on six-week intercept: F(1, 426) = 214.96, p < .001, d = 1.99 (responders: M = 27.83 words, SD = 8.86; nonresponders: M = 13.65 words, SD = 5.22); and slope: F(1, 426) = 116.77, p < .001, d = 1.52, per week (responders: M = 1.99 words/week, SD = 0.76; nonresponders: M = 1.09 words/week, SD = 0.40).

The 232 nonresponders were then given a pretest battery that included measures of phonemic decoding efficiency, sight-word reading efficiency, untimed decoding skill, and untimed word identification skill. Multiple measures of the same construct were obtained for the purpose of creating a latent word-level reading factor, which is preferable to using individual measures of word reading because measurement error is eliminated in the former but not in the latter (McArdle, 2009). Further, word reading (comprising timed and untimed tests of real and nonsense words) was chosen as the outcome of interest because the primary focus of our instructional materials was on improving word reading. Three students moved during the pretesting period, and one was removed due to lack of verbal assent.

**Assigning At-Risk Students to Tier 2 Instruction**

Students identified as unresponsive to Tier 1 were randomly assigned to one of two conditions: continued Tier 1 (n = 80, 35.09%) or Tier 2 (i.e., supplemental small-group tutoring; n = 148, 64.91%). A larger proportion of students was assigned to Tier 2 to allow for later assignment of students who were unresponsive to Tier 2 either to continue in Tier 2 or to receive Tier 3 (i.e., one-on-one tutoring). Within the Tier 2 condition, we placed students (in the same school) into tutoring groups of three or four according to their PM slopes and intercepts. This phase of tutoring consisted of seven weeks of instruction, and the progress of students in the Tier 2 tutoring condition and the Tier 1 control condition was monitored on a weekly basis. Seven students in the Tier 2 condition moved during the seven weeks of tutoring, one student was removed at the request of the parent, one was removed due to a teacher request, and one was removed due to behavior problems.

**Identifying and Randomly Assigning Students Who Were Unresponsive to Tier 2 Tutoring**

PM data (consisting of weekly one-minute WIF assessments) from the first seven weeks of Tier 2 tutoring were used to estimate tutored students’ slopes and intercepts. Again, students were rank ordered on PM growth and PM level; the bottom 45 students (20.36%) were identified as nonresponsive to small-group tutoring. Responders (n = 89) to the first seven weeks of Tier 2 tutoring
outperformed nonresponders ($n = 45$) on intercept: $F(1, 32) = 127.40, p < .001, d = 2.07$ (responders: $M = 27.25$ words, $SD = 8.26$; nonresponders: $M = 12.01$ words, $SD = 5.12$); and slope: $F(1, 132) = 9.99, p < .001, d = 1.52$ (responders: $M = 1.06$ words/week, $SD = 0.31$; nonresponders: $M = 0.48$ words/week, $SD = 0.15$).

Unresponsive students were randomly assigned to continue receiving Tier 2 tutoring ($n = 21$) or move to Tier 3 tutoring ($n = 24$). We defined Tier 3 as one-on-one daily instruction. Before beginning Tier 3 tutoring, students were given a short screener to determine point of reentry into the tutoring curriculum. A reevaluation

### TABLE 1
Latent Class Analysis for Cohorts 1 and 2 to Identify the Initial Risk Pool of Students ($N = 628$)

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean group performance</th>
<th>Model evaluation criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RLN</td>
<td>WIF1</td>
</tr>
<tr>
<td><strong>Cohort 1 ($n = 330$)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-class solution</td>
<td>7,683.10</td>
<td>525.53,</td>
</tr>
<tr>
<td>Class 1 ($n = 282$)</td>
<td>39.08</td>
<td>14.67</td>
</tr>
<tr>
<td>Class 2 ($n = 48$)</td>
<td>53.14</td>
<td>50.73</td>
</tr>
<tr>
<td>Three-class solution</td>
<td>7,488.75</td>
<td>208.55,</td>
</tr>
<tr>
<td>Class 1 ($n = 223$)</td>
<td>37.76</td>
<td>12.78</td>
</tr>
<tr>
<td>Class 2 ($n = 72$)</td>
<td>51.21</td>
<td>34.31</td>
</tr>
<tr>
<td>Class 3 ($n = 35$)</td>
<td>53.21</td>
<td>60.39</td>
</tr>
<tr>
<td>Four-class solution</td>
<td>7,532.92</td>
<td>115.97,</td>
</tr>
<tr>
<td>Class 1 ($n = 203$)</td>
<td>36.08</td>
<td>11.95</td>
</tr>
<tr>
<td>Class 2 ($n = 74$)</td>
<td>50.21</td>
<td>28.29</td>
</tr>
<tr>
<td>Class 3 ($n = 39$)</td>
<td>51.48</td>
<td>48.35</td>
</tr>
<tr>
<td>Class 4 ($n = 14$)</td>
<td>56.69</td>
<td>70.01</td>
</tr>
<tr>
<td><strong>Cohort 2 ($n = 298$)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-class solution</td>
<td>7,007.01</td>
<td>447.24,</td>
</tr>
<tr>
<td>Class 1 ($n = 256$)</td>
<td>42.51</td>
<td>15.17</td>
</tr>
<tr>
<td>Class 2 ($n = 42$)</td>
<td>54.23</td>
<td>50.78</td>
</tr>
<tr>
<td>Three-class solution</td>
<td>6,851.23</td>
<td>171.06,</td>
</tr>
<tr>
<td>Class 1 ($n = 214$)</td>
<td>41.96</td>
<td>14.11</td>
</tr>
<tr>
<td>Class 2 ($n = 56$)</td>
<td>51.82</td>
<td>39.66</td>
</tr>
<tr>
<td>Class 3 ($n = 28$)</td>
<td>59.16</td>
<td>67.77</td>
</tr>
<tr>
<td>Four-class solution</td>
<td>6,914.31</td>
<td>153.32,</td>
</tr>
<tr>
<td>Class 1 ($n = 169$)</td>
<td>38.91</td>
<td>10.39</td>
</tr>
<tr>
<td>Class 2 ($n = 84$)</td>
<td>49.26</td>
<td>23.77</td>
</tr>
<tr>
<td>Class 3 ($n = 32$)</td>
<td>51.73</td>
<td>42.51</td>
</tr>
<tr>
<td>Class 4 ($n = 13$)</td>
<td>59.21</td>
<td>68.23</td>
</tr>
</tbody>
</table>

Note. BIC = Bayesian information criteria; $-2\text{(log-likelihood value of the model)} + (\text{number of model parameters} \times \text{natural log of the sample size})$. LMR LRT = Lo–Mendell–Rubin likelihood ratio test; from “Testing the Number of Components in a Normal Mixture,” by Y. Lo, N. Mendell, & D.B. Rubin, 2001, *Biometrika*, 88(3), 774. RLN = rapid letter naming (letters/min). WIF = word identification fluency (words/min). Lower scores represent better fitting models. LMR LRT compares the estimated model with a model with one less class than the estimated model. A significant $p$-value indicates that the model with one less class is rejected in favor of the estimated model. Entropy is a summary measure of the probability of membership in the most likely class for each individual. There are no specific guidelines for interpreting entropy, but possible values range from 0 to 1.0, and values closer to 1.0 represent better classification.
of tutoring level was necessary because these unresponsive participants had not maintained pace academically with their Tier 2 small-group peers. This second phase of tutoring also lasted seven weeks, and all students continued to receive weekly PM. Two students moved during this phase of the study.

**Posttesting**

The posttest battery (comprising the same assessments used in the pretesting battery: phonemic decoding efficiency, sight-word reading efficiency, untimed decoding skill, and untimed word identification skill) was administered to 395 (90.62%) of students who had been identified with potential risk by the initial screening. Forty-two students moved or were removed at some point during the year. Five students were omitted from further analyses due to incomplete data. Comparisons between students who completed the posttest battery (n = 395) and those who did not (n = 47) revealed no significant differences for gender: \( \chi^2(2, N = 437) = 5.48, p = .065 \); race: \( \chi^2(6, N = 436) = 3.00, p = .808 \); free/reduced lunch status: \( \chi^2(1, N = 436) = 1.54, p = .215 \); and WIF screening scores: \( F(1, 433) = 0.22, p = .637 \); or RLN screening scores: \( F(1, 433) = 1.81, p = .179 \).

**Final Sample Description for Efficacy Analysis**

Because our primary interest is in the efficacy of multi-tiered instruction for students who are at risk for later reading difficulty, our final sample comprised only those students who were unresponsive to Tier 1 instruction (n = 212). Table 2 contains demographic and screening score information for this group of students. Males (51.89%) comprised just over half of the sample, as did those who qualified for a free or reduced-price lunch (66.04%). The ethnicity makeup was primarily African American (46.70%) and Caucasian (37.74%). On average, Tier 1 nonresponders were naming 35.14 letters/minute and 8.14 words/minute. The fact that no students received a score of 0 on the letter-naming task and only five students read no words on the timed word list indicates that nearly all students in our sample had at least minimal foundational word reading skills.

Follow-up assessments were administered in the spring of grades 2 and 3. The prevention analysis was conducted on only the 295 students with complete data through third grade, which represents an attrition rate of 25.32% for our sample. The measures used at follow-up were the same as those used in pre- and posttesting battery (phonemic decoding efficiency, sight-word reading efficiency, untimed decoding skill, and untimed word identification skill). Individual testing sessions were scheduled so trained research assistants could administer the battery during school hours at times that teachers specified as being least disruptive to academic learning.

**Intervention**

**Theoretical Orientation**

The instructional focus of the activities included in the supplemental, remedial tutoring program were letter-sound correspondence, sight-word recognition, phonemic awareness, decoding, spelling, and reading fluency (Foorman & Torgesen, 2001; Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001). These activities directly address the first three of the five essential components of effective reading instruction according to the National Reading Panel (National Institute of Child Health and Human Development, 2000): phonemic awareness, phonics, and fluency. Strengthening of foundational word reading skills of young students is critical because doing so has been shown to improve not only word reading but also reading comprehension (Coyne, Kame‘enui, Simmons, & Harn, 2004; Ehri, Nunes, Stahl, & Willows, 2001; Vadsay, Sanders, & Abbott, 2008).

These activities that are foundational to reading development were included in the RTI models described in the introduction; in addition, two studies also included activities related to comprehension and vocabulary (Case et al., 2010; Wanzek & Vaughn, 2008), one included a writing activity (Vadsay et al., 2000), and one included instruction in speech articulation (Kerins et al., 2010). Before describing the remedial tutoring program.

---

**TABLE 2**

**Demographic Information and Screening Scores for Students Unresponsive to General Classroom Instruction (N = 212)**

<table>
<thead>
<tr>
<th>Demographic information</th>
<th>f</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>110</td>
<td>51.89</td>
</tr>
<tr>
<td>F/R lunch</td>
<td>140</td>
<td>66.04</td>
</tr>
<tr>
<td>African American</td>
<td>99</td>
<td>46.70</td>
</tr>
<tr>
<td>Asian</td>
<td>2</td>
<td>0.94</td>
</tr>
<tr>
<td>Caucasian</td>
<td>80</td>
<td>37.74</td>
</tr>
<tr>
<td>Hispanic</td>
<td>17</td>
<td>8.02</td>
</tr>
<tr>
<td>Other ethnicity</td>
<td>14</td>
<td>6.61</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Screening scores</th>
<th>Mean (SD)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>RLN (letters/min)</td>
<td>35.14 (10.76)</td>
<td>4–68</td>
</tr>
<tr>
<td>WIF (words/min)</td>
<td>8.14 (4.77)</td>
<td>0–23.5</td>
</tr>
</tbody>
</table>

Note. F/R Lunch = free/reduced lunch status. RLN = rapid letter naming. WIF = word identification fluency. n = 210 for Male variable. n = 211 for RLN and WIF.
used in this study, we should make clear that we endorse it only as a *supplement* not a substitute for a literacy-rich comprehensive reading curriculum made up of evidence-based practices implemented by highly qualified classroom teachers in Tier 1.

The primary difference in our tiers of instruction was intensity, which we defined as decreased student to instructor ratio and increased frequency of instruction (Foorman & Torgesen, 2001). These elements have been used in other studies seeking to improve the word reading of students who have not responded well to prior instruction (McMaster et al., 2005; O'Connor et al., 2005; Vadasy, Sanders, Peyton, & Jenkins, 2002; Vellutino et al., 1996; Wanzek & Vaughn, 2008). Smaller instructional groupings and additional opportunities to access instructional material are likely to benefit students who may need a high-engagement setting (one with few opportunities for off-task behavior) and additional exposures to the material to be successful. Therefore, Tier 2 differed from Tier 1 in that the former was implemented in a small-group rather than whole-class format, and tutoring was provided as a supplement to classroom instruction, thus increasing the frequency of instruction. Tier 3 was more intensive than Tier 2 in that the former was provided in a one-on-one format and was implemented daily rather than three times a week.

Training

Graduate research assistants were provided five weeks of tutoring training, which began with a two-hour introduction to the standard tutoring program. Then, each research assistant was required to practice the tutoring program for 17 hours. Finally, each research assistant completed a mock tutoring session with the trainer, who addressed all discrepancies as the session was conducted. Across both cohorts, 25 research assistants provided tutoring.

Tier 2

Trained research assistants provided small-group tutoring three times per week in 45-minute sessions. Homogeneous small groups were formed to provide tutored students with the highest level of instruction possible that was most likely to benefit the weakest reader in the group. Before beginning instruction, students within each small group were administered a short screenser to determine point of entry into the tutoring curriculum. The lowest estimated entry point within each group served as the beginning point for group instruction. Tutoring was conducted in a supplementary pull-out context and included activities that have been found to have a positive effect on the performance of struggling readers and students with reading disabilities (e.g., Blachman, Tangel, Ball, Black, & McGraw, 1999; Hatcher, Hulme, & Ellis, 1994; Torgesen, 2000).

Tutors followed scripted protocols that standardized implementation of instruction to address the following skills: letter–sound correspondence, sight-word recognition, phonemic awareness, decoding, and reading fluency. Each of the eight instructional activities corresponded to a series of eight leveled reading books. For the portion of the book that was covered in a particular lesson, each word fit into one of three categories: sight word, decodable word, or story word. The sight-word category represents the high-frequency words, the decodable word category represents words that can be sounded out using the sounds students had been taught up to the present lesson in the program, and the story word category represents all the words that are not considered either sight words or decodable words.

Each tutoring session began with 10 minutes of letter–sound correspondence training and sight-word instruction. For letters, each lesson included a mix of new and previously taught letter–sound correspondences, totaling approximately 12 per lesson. New letter–sound correspondences were first modeled by the tutor, after which students were instructed to repeat the pronunciation. After all letter sounds had been introduced, students were prompted to recall the sounds associated with letters chorally. Finally, each student was asked to recall each sound individually. For incorrect answers, the tutor corrected the student and later revisited all the previously missed letter–sound correspondences.

For sight words, a similar procedure was used. The tutor introduced each sight word by presenting a flashcard with the word typed on it and then pronouncing and spelling the word. Students were instructed to repeat the pronunciation and spelling. After all words were introduced, students were prompted to read the sight words chorally. Finally, each student was asked to read each word individually. For incorrect answers, the tutor corrected the student and later revisited all the previously missed words. If a student did not respond with the correct answer after revisiting it, the tutor retaught that word to the whole group during the next lesson.

The second activity was the introduction of story words. The tutor presented a flashcard with the story word typed on it and then pronounced it for the students. Students were asked to repeat the pronunciation. This activity lasted five minutes.

The third activity addressed phonemic awareness and decoding ability. Using the decodable words of the lesson, tutors modeled how to tap out each sound in the word using a finger or pencil. They asked students to imitate the tapping. After all words had been tapped, tutors modeled how to blend the sounds in the words. Students repeated the blending, blended chorally as a
group, and then blended individually. Tutors corrected errors and revisited missed words. Finally, tutors provided the students with paper and pencils and asked students to spell the decodable words in the lesson. Phonological awareness and decoding lasted 10 minutes.

In the fourth activity, tutor and students reviewed the sight words, story words, and decodable words that the students had learned during that lesson. Each student was given one sentence from the day’s story. Students were asked to search for and highlight each sight word called out by the tutor until all sight words on all sentence strips were highlighted. Students who had difficulty in finding a word were allowed to see the flashcard as an aid in finding the word. This process was repeated for story words and decodable words. The sentence strip activity lasted seven minutes.

The last treatment component was text reading and fluency. The activity started with the tutor and students reading the previous lesson’s portion of the book together. Then, the tutor read a new portion of the book and asked students to repeat the reading in a line-by-line fashion. Afterward, the tutor and students chorally read the new portion. Finally, each student was given two opportunities to read the new portion individually and was encouraged to read more quickly on the second attempt. Throughout the lesson, tutors offered points for good behavior and effort, which could be redeemed for small prizes.

**Tier 3**

Trained research assistants provided 30 minutes of one-on-one tutoring five days a week. Lessons followed the same lesson guides as the small-group tutoring except that the tutor took the place of the other group members during choral response.

**Fidelity**

Each lesson was audiotaped. A fidelity-of-implementation checklist was created from the tutoring script, and a random sample of 20% of the lessons revealed tutors accurately implemented the 64 steps with an average accuracy of 94.04%.

**Pre- and Posttest Measures**

**Sight-Word Reading Efficiency**

The Test of Sight Word Reading Efficiency (Torgesen, Wagner, & Rashotte, 1997) is a norm-referenced measure of sight-word reading accuracy and fluency. For this test, students are presented with a list of 104 words ordered in ascending difficulty and asked to read aloud as many as possible in 45 seconds. Test–retest reliability exceeded .80 for the sample.

**Phonemic Decoding Efficiency**

The Test of Phonemic Decoding Efficiency (Torgesen et al., 1997) is a norm-referenced measure of decoding accuracy and fluency. For this test, students are presented with a list of 63 nonsense words ordered in ascending difficulty and asked to read (decode) aloud as many as possible in 45 seconds. Test–retest reliability exceeded .80 for the sample.

**Untimed Decoding Skill**

The Woodcock Reading Mastery Test–Revised—Normative Update: Word Attack (Woodcock, 1998), a norm-referenced test, evaluates students’ ability to pronounce pseudowords presented in list form. It contains 45 nonsense words ordered from easiest to most difficult. Students are asked to read (decode) the words aloud, one at a time. The developer-recommended basal and ceiling rules were applied to minimize boredom and frustration. Split-half reliability exceeded .90 for the sample.

**Untimed Word Identification Skill**

The Woodcock Reading Mastery Test–Revised—Normative Update: Word Identification (Woodcock, 1998), a norm-referenced test, asks students to read single words in list form. It consists of 100 words ordered in difficulty. Students are asked to read (decode) the words aloud, one at a time. Again, the developer-recommended basal and ceiling rules were applied to minimize boredom and frustration. Split-half reliability exceeded .90 for the sample.

**Data Analysis**

Data analysis was conducted in four steps. First, students were classified into groups according to their assessment, treatment, and response status throughout the school year. This allowed us to form the appropriate contrasts for the tutoring efficacy analysis (research questions 1 and 2) and the prevention analysis (research question 3). Then, contrast codes were created to compare at-risk students who did and did not receive supplemental tutoring (research question 1) and Tier 2 nonresponders who did and did not receive Tier 3 tutoring (research question 2). Our third step was to run a multilevel model to evaluate research questions 1 and 2 while also accounting for dependency in the data due to small-group and classroom membership. Our final step was to calculate and report proportions of students in each group who had word reading scores in the normal range (>30th percentile) at the end of grades 1–3 (research question 3).

**Classifying Students Into Groups**

First, to facilitate the analyses, we classified students into distinct groups according to their assessment, treatment, and response status throughout the year (see Figure 1).
Group 1 (n = 218) comprises students who were initially identified as not at risk for reading difficulty after stage 1 universal screening; this group was excluded from further study participation and was therefore excluded from analyses. The remaining groups include only students who completed the posttest battery. Group 2 comprises students who were initially identified as potentially at-risk during stage 1 universal screening, but whose six weeks of PM disconfirmed risk by response to general education (n = 183). These students were administered only the posttest battery, not the pretest battery. Groups 2–4 (n = 212) were subsets of students who were identified as unresponsive to Tier 1, all of whom were pretested, were posttested, and received weekly PM with WIF. Of these students, the 78 who were randomly assigned to remain in Tier 1 without tutoring comprised group 3.

After the initial wave of seven weeks of small-group tutoring (i.e., Tier 2), we used PM data to determine subsets of students who were (un)responsive. Those who were responsive remained in Tier 2 for an additional seven weeks and were designated as group 4 (n = 89). Those who were unresponsive to Tier 2 tutoring were randomly assigned to remain in Tier 2 (group 5, n = 21) to serve as a control group for Tier 3 or to receive daily one-on-one (Tier 3) tutoring (group 6, n = 24). A description of each group can be found in Table 3.

**TABLE 3**

Description of Groups and Purpose in Analysis

<table>
<thead>
<tr>
<th>Group</th>
<th>Description</th>
<th>Purpose in efficacy analyses (research questions 1 and 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Identified as not at-risk by initial screening</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>Identified as at-risk by initial screening</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Risk disconfirmed by responsiveness to Tier 1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Identified as at-risk by initial screening</td>
<td>Used as a control group for at-risk students who received some type of supplemental intervention</td>
</tr>
<tr>
<td></td>
<td>Risk confirmed by unresponsiveness to Tier 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Randomly assigned to remain in Tier 1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Identified as at-risk by initial screening</td>
<td>Combined with groups 5 and 6 to assess efficacy of supplemental intervention</td>
</tr>
<tr>
<td></td>
<td>Risk confirmed by unresponsiveness to Tier 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Randomly assigned to Tier 2 for 7 weeks</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Responsive to Tier 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Remained in Tier 2 for additional 7 weeks</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Identified as at-risk by initial screening</td>
<td>Combined with groups 4 and 6 to assess efficacy of supplemental intervention and used as a control group for the efficacy of Tier 3</td>
</tr>
<tr>
<td></td>
<td>Risk confirmed by unresponsiveness to Tier 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Randomly assigned to Tier 2 for 7 weeks</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unresponsive to Tier 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Randomly assigned to additional 7 weeks of Tier 2</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Identified as at-risk by initial screening</td>
<td>Combined with groups 4 and 5 to assess efficacy of supplemental intervention and used to assess the efficacy of Tier 3</td>
</tr>
<tr>
<td></td>
<td>Risk confirmed by unresponsiveness to Tier 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Randomly assigned to Tier 2 for 7 weeks</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unresponsive to Tier 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Randomly assigned to Tier 3 for 7 weeks</td>
<td></td>
</tr>
</tbody>
</table>

**Weighted Contrast Codes**

Treatment effects were analyzed via weighted contrast codes in a multilevel regression model. Student membership in schools, classrooms, and small groups was accounted for in the model to control for dependency in the data. We created three (orthogonal) weighted contrast codes to assess treatment effects for nonresponders to Tier 1; students in group 2 were excluded from the efficacy analysis because they showed low risk for reading difficulty after six weeks of PM during Tier 1.

The first contrast code, W456_3, was created to compare the performance of at-risk students who received supplemental tutoring (groups 4–6) and those who received only Tier 1 instruction (group 3). The second contrast code, W6_5, was created to compare the performance of students who, after being initially identified as unresponsive to Tier 2, received Tier 3 (group 6) versus continued Tier 2 instruction (group 5). Finally, we created W4_56 to complete the set of orthogonal contrasts, but this comparison of end-of-year performance between students identified as responders to Tier 2 (group 4) and nonresponders to Tier 2 (groups 5 and 6) was not of theoretical interest. Weighted contrast codes were used because the groups of interest for comparison had unequal sample sizes.
Treatment Effects on Latent Change

Before estimating the multilevel regression model, we employed a latent change score analysis to calculate the change in word reading ability from pretest to posttest. Figure 2 represents the general latent change model proposed by McArdle (2009). Latent change models alleviate many of the measurement error problems associated with simple change scores (i.e., unreliability of the change score, elevated variance estimates, regression to the mean effects) by providing variance estimates of the latent variable at each time point that are free of error, and therefore the variance in a latent change score is not confounded by errors of measurement (see Little, Bovaird, & Slegers, 2006). Thus, multivariate latent change scores avoid problems associated with unreliable difference scores’ regression to the mean (McArdle & Nesselroade, 1994).

In this analysis, pretest measures of word identification, word attack, sight-word efficiency, and phonemic decoding efficiency were allowed to load onto a pretest factor variable ($f_1$). Similarly, posttest measures of the same assessments were allowed to load onto a posttest factor variable ($f_2$). The latent change from pretest to posttest is represented by $\Delta f$. Latent change analysis more accurately represents change in word reading ability than any one of the measures alone. We extracted estimates of $f_1$ and $\Delta f$ from the model, standardized them to have a $M$ of 0 and $SD$ of 1, and entered them into a multilevel regression model.

Multilevel modeling accounts for dependency in data that exists because of shared contexts that prevent individual cases to be truly independent from one another. Independence is an underlying assumption of multiple regression models, so it is necessary to account for dependency by adding random effects for contexts (or levels, as in multilevel modeling) to obtain accurate standard errors for coefficient estimates (Raudenbush & Bryk, 2002). Although the same sampling procedures were used in both years of the study, we included cohort in the model to control for any unexpected differences between the two cohorts.

After first determining that there were no significant cohort × treatment interactions, our proposed model for assessing the efficacy of supplemental tutoring is represented in equation 1:

$$z\Delta f_{ijkmn} = B_0 + B_1 \text{Cohort}_{ijkmn} + B_2 z f_{1ijkmn} + B_3 W_{456} - 3_{ijkmn} + B_4 W_{6} - 5_{ijkmn} + B_5 W_{456} - 5_{ijkmn} + u_{0j} + u_{0k} + u_{0m} + e_{ijkmn},$$

where $u_{0j}$ is the random effect associated with small groups in the first phase of tutoring, $u_{0k}$ is the random effect associated with small groups in the second phase of tutoring, $u_{0m}$ is the random effect associated with classrooms, $u_{0n}$ is the random effect associated with schools, and $e_{ijkmn}$ is the residual (or error) term.

We did not allow treatment effects to vary across small groups, classrooms, or schools because we expected the differences between the various treatment comparisons to be the same across those contexts.

Prevention

To assess the multitiered tutoring system’s capacity to prevent later reading difficulties for students who show early risk for such outcomes, we calculated the proportion of at-risk students who received tiered tutoring and achieved a score on each of the four word

FIGURE 2
Latent Change Model of Word Reading Ability

Note. 1 = pretest. 2 = posttest. PDE = phonemic decoding efficiency. SWE = sight-word efficiency. WA = word attack. WID = word identification.
reading measures above the 30th percentile. We chose the 30th percentile criterion because past research on RTI has used this cut score (Mathes et al., 2005; McMaster et al., 2005; Torgesen, 2000), and we required adequate performance on each test rather than the mean of the tests because we consider decoding (word attack and phonemic decoding efficiency), word identification (word identification and sight-word efficiency), accuracy (word attack and word identification), and fluency (phonemic decoding efficiency and sight-word efficiency) to all be critical dimensions of proficient word reading. Therefore our criterion required students to display adequate achievement on each dimension so relative weaknesses on any of the dimensions could not be masked by relative strengths on the others.

Longitudinal data are necessary to answer questions of prevention, but unfortunately missing data due to participant attrition increases with each subsequent year of data collection. For the subset of students who had follow-up data through grade 3, we report the proportion of students scoring above the 30th percentile on each word reading measure in the spring of each grade. Specifically, we were interested in the difference in normal achievement between at-risk students in Tier 2 (groups 4–6) compared with those in Tier 1 (group 3), with a closer examination of those students who responded versus did not respond to Tier 2 (group 4 and groups 5 and 6, respectively).

Results

For each group, pre- and posttest scores for the measures that comprised the latent word reading factor are listed in Table 4. All groups made gains from pretest to posttest on all measures, but some gains were higher than others. Correlations (among the pre- and posttest measures listed in Table 5) were moderate, ranging from .55 to .85 and from .63 to .88, respectively. Measures were correlated with each other approximately the same at pretest as they were at posttest. The biggest discrepancy was that timed decoding was slightly more highly correlated with the other measures at posttest than at pretest.

Research questions 1 and 2 were answered using estimates from the same multilevel model (equation 1). Once again, multilevel modeling was used to account for possible dependency in the outcome measure as a result of small-group or classroom differences. Research question 3 was answered in a more descriptive way, proportion counts of students achieving normal range word reading performance were reported and compared across tiers and responder groups.

Research Question 1: Intervention Efficacy for Nonresponders to Tier 1

Our first research question was, Is 14 weeks of scripted, supplemental tutoring (Tier 2 and/or Tier 3) effective for students identified as unresponsive to Tier 1 instruction? Fit statistics for the latent change model depicted in
Efficacy of a First-Grade Responsiveness-to-Intervention Prevention Model for Struggling Readers

Figure 2 indicated the model combining cohorts was a good fit for the data: $\chi^2(22, N = 212) = 168, p > .05$; comparative fit index = .978, Tucker–Lewis index = .952; root mean square error = .001; standardized root mean square residual = .059 (see Hu & Bentler, 1998). Therefore, we entered standardized estimates of $z_f$ and $z_D$ from the latent change model into the model in equation 1 to assess treatment effects. Residuals were examined for potential outliers. Five observations had externally studentized residuals >2.5 that were also visually distinct from the remaining sample on a number of graphical displays; further, these observations had relatively high influence on the model (difference in fit standardized value; Cohen, Cohen, West, & Aiken, 2003). The outliers were omitted, and the model was reestimated. Fixed and random effects are presented in Table 6.

From the significant coefficient associated with the $W_{456_3}$ variable ($\hat{B}_3 = 0.001, t = 1.99, p = .048$), we conclude that for students who were deemed at risk for reading difficulties because of their nonresponse to Tier 1 instruction, supplemental reading tutoring was beneficial. Students who received tutoring (groups 4–6 combined representing Tiers 2 and 3), on average, had greater change scores than did students who received reading instruction only in their classrooms (group 3 representing Tier 1), controlling for pretest word reading ability.

Using the student-level effect size equation found in the What Works Clearinghouse Procedures and Standards Handbook (What Works Clearinghouse, 2008), we found the standardized mean difference across tutoring phases. $W_{456_3}$ = weighted effect for responders (group 4) vs. nonresponders (groups 5 and 6) during first phase of tutoring. $W_{456_3}$ = weighted effect for treatment (groups 4–6) vs. control (group 3) across tutoring phases. $W_{6_5}$ = weighted effect for one-on-one tutoring (group 6) vs. small-group tutoring (group 5) during second phase of tutoring. $z_f = \text{latent pretest } z$-score.

where $X$ represents the mean change of the students in the respective groups, and $S$ represents the standard deviation of change among the students in the respective groups.

We did not correct for clustering because conditional variance at each level was either nonexistent or nonsignificant in the latent change model (see random effects in Table 6). Furthermore, assignment to treatment (groups) was determined at the student level based on each student’s response to their current tier of instruction.

**Research Question 2: Intervention Efficacy for Nonresponders to Tier 2**

Our second research question was, Is Tier 3 instruction (delivered five days a week in a one-on-one tutoring format) more effective than continued Tier 2 (delivered three days a week in a small-group tutoring format) for those students who fail to respond to Tier 2? Of the students who did not respond to the initial phase of Tier 2, there was no difference in change scores between those who received Tier 3 (group 6) and those who received a second wave of Tier 2 (group 5: $\hat{B}_4 = -0.009, t = -1.72, p = .086$).

One caution regarding this finding is that although students were randomly assigned to Tier 2 and Tier 3...
instruction, WIF scores from the week just prior to assignment were not equivalent between the groups: \( F(1, 41) = 4.69, \ p = .036 \) (two students were missing WIF scores for that particular week, hence the 42 degrees of freedom). The group assigned to Tier 2 was reading an average of 13.55 words/minute, whereas the group assigned to Tier 3 was reading an average of only 9.27 words/minute. However, their difference on \( f1 \) was not significantly different: \( F(1, 43) = 2.92, \ p = .09 \); and a model that includes both \( f1 \) and the prior WIF score did not result in a significant difference between the two groups on \( \Delta f \) (\( p = .38 \)). Thus, we have no evidence that the nonsignificant Tier 3 results are explained by differences in prior word reading ability.

**Research Question 3: Prevention**

Our third research question was, What proportion of at-risk students who receive Tier 2 instruction achieves reading performance in the normal range in grades 1–3? One purpose of this study was to determine whether multitiered supplemental tutoring would be able to prevent later reading difficulties for students who were considered at risk for reading difficulties based on screening and PM data (during Tier 1 instruction). Proportions of at-risk students in Tiers 1 and 2 who achieved adequate word reading levels are presented in Table 7.

At the end of grade 1, slightly more students in Tier 2 (59%) scored in the average range on word reading than did students in Tier 1 (53%); this corroborates our efficacy analysis above (effect size = .19). Over time, though, both groups showed a decline in the proportion of average readers, and the two groups had the same proportion of students who were reading in the average range. In second grade, average readers comprised 46% of both groups. Third-grade proportions were lower: 39% for Tier 1 and 40% for Tier 2. Proportions of responders compared with nonresponders to Tier 2 who were reading in the average range differed dramatically across the grades. By third grade, 53% of the responders and 12% of the nonresponders were reading in the average range. Thus, multitiered supplemental tutoring was not associated in the long term with a higher proportion of at-risk students achieving average range word reading than regular classroom instruction; even response to Tier 2 instruction left 47% of the students failing to reach average range achievement in word reading in third grade.

**Discussion**

Results indicate that students who are at risk for reading difficulty because of unresponsiveness to Tier 1 classroom instruction benefit from supplemental reading tutoring that is based on a standard program approach compared with at-risk students who continue in their Tier 1 classroom without supplemental reading tutoring. This corroborates Mathes et al.'s (2005) study, in which Tier 2 tutoring effects were evident on standardized measures of timed and untimed word reading. In the present study, tutoring was effective in boosting the development of at-risk first-grade readers with a modest effect size of .19.

Participation in Tier 2 instruction, though, did not necessarily prevent future reading difficulties. At the end of grade 1, 41% of at-risk students who participated in supplemental tutoring failed to score in the normal, or average, range (compared with 47% of at-risk students receiving only Tier 1). These findings are somewhat consistent with reports from other multitiered tutoring systems. Our figures are closest to those reported by McMaster et al. (2005). In their study, 45% of at-risk students who participated in Tier 2 intervention did not reach the normal range of performance on decoding and/or word identification.

Lower proportions of below-average reading were reported by Vadasy et al. (2000) and Mathes et al. (2005). Here, a difference of criterion should be noted: In the Mathes et al. study, students were required to reach normal range achievement on an average of two measures (decoding and word identification accuracy) instead of both. One difficulty in comparing these proportions of below-average range performance across studies is differences in identification of the risk pool. As we stated, our proportions were closest to McMaster et al.'s, and it is likely no coincidence that risk identification procedures were closely aligned for both studies (poor performance on screening measure and dual discrepancy for PM slope and intercept). The risk identification procedures for Vadasy et al. and Mathes et al. did not include monitoring student progress over time and therefore may have included a number of false positives in their risk pools; this may explain their lower rates of below-average reading performance.

**TABLE 7**

<table>
<thead>
<tr>
<th>Group</th>
<th>( n )</th>
<th>Grade 1</th>
<th>Grade 2</th>
<th>Grade 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tier 1 (group 3)</td>
<td>57</td>
<td>.53</td>
<td>.46</td>
<td>.39</td>
</tr>
<tr>
<td>Tier 2 (groups 4–6)</td>
<td>103</td>
<td>.59</td>
<td>.46</td>
<td>.40</td>
</tr>
<tr>
<td>Tier 2 responders</td>
<td>70</td>
<td>.71</td>
<td>.61</td>
<td>.53</td>
</tr>
<tr>
<td>(group 4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tier 2 nonresponders</td>
<td>33</td>
<td>.33</td>
<td>.12</td>
<td>.12</td>
</tr>
<tr>
<td>(groups 5 and 6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Numbers in parentheses represent the percentage of students from the original group who had complete data through third grade.
Our longitudinal data allow us to examine the proportions of at-risk students who achieve reading scores in the average range over time. Compared with grade 1, the proportions are lower for grades 2 and 3. At the end of third grade, 60% of the at-risk students who participated in supplementary tutoring in the form of Tier 2 and/or Tier 3 instruction failed to attain reading scores in the average range. Even response to Tier 2 instruction did not prevent later reading difficulties for many of the students; 47% of the responders still failed to achieve word reading scores above the 30th percentile.

We believe that early prevention is important because it potentially affords students identified as at risk for future reading problems greater potential to benefit from regular classroom instruction. Additionally, it increases the likelihood that students will gain content knowledge from text as they read (Gough, 1996). To the extent that results from our 14 weeks of first-grade multitiered supplemental tutoring are representative of what can be expected in schools, we conclude that the intervention was not sufficient to prevent the later reading difficulties of students who are at risk for such outcomes. We infer that the supplemental preventive programs associated with RTI may need to span multiple years to accomplish the preventive intent. As shown in previous work, multiple years (Vellutino et al., 1996) or multiple waves (O’Connor et al., 2005; Vaughn, Linan-Thompson, & Hickman, 2003) of intervention may be required to help these students perform within the average range. Of course, the level of response required to ensure students thrive when they return to Tier 1 is an empirical question in need of further investigation.

In terms of addressing the deficits of students who do not respond to Tier 2 supplemental tutoring, our results indicate that providing increases in the intensity of Tier 2 intervention, which we operationalized as more sessions per week with one-on-one delivery, was not a viable solution for improving the reading skills of students found to be unresponsive to Tier 2 intervention. This is in line with previous research suggesting that higher doses of the same Tier 2 tutoring (Vadasy et al., 2002; Wanzek & Vaughn, 2008) or smaller teacher-to-student ratios (Iversen, Tunmer, & Chapman, 2005; Vaughn, Linan-Thompson, Kouzekanani, et al., 2003) have limited effects over typical small-group Tier 2 practices on the reading skills of chronically struggling students. This study extends current knowledge by demonstrating that the combination of higher doses and decreased teacher-to-student ratios does not, in and of itself, produce increased reading scores over lower doses and larger groupings of students.

The deficits of students who require Tier 3 intervention may be better addressed by an individualized, or problem-solving, approach to RTI in which intervention and assessment are specially designed to meet the needs of each individual student (D. Fuchs & Fuchs, 2006), akin to individualized education programs found traditionally in special education. More recently, individualized instruction at Tier 3 has been dubbed experimental teaching or data-based program modification (D. Fuchs, Fuchs, & Compton, 2012) to emphasize the roles of data-based decision making and practitioner expertise. In this type of intervention, special educators would initially start with a standard protocol program and then make modifications (based on their expert judgment) when PM data suggest that the current instruction is not meeting the instructional needs of the student.

We acknowledge the potential value in individualized approach, yet we attempted the standard protocol approach because of the experimental control that can be achieved with a standard treatment rather than individualized instructional programming. In addition, standard protocol approaches tend to be more efficient on the whole due to standardized implementation of research-based interventions, fixed duration of intervention, and set assessment occasions for all students who are classified as at risk for later reading difficulties. Yet, to the extent that our Tier 3 results are generalizable, we have shown the limits of an intensive short-term standard protocol approach for chronically struggling readers. Meeting the instructional needs of students who fail to respond to standard protocol supplemental tutoring may require the more resource-intensive individualized approach (problem-solving or experimental teaching).

Two other possibilities exist for intensifying instruction that we did not address in this study: increasing duration and instructor expertise. O’Connor et al. (2005) extended Tier 3 instruction across multiple grades and showed that students who received multitiered instruction had substantially greater scores than did at-risk students who did not receive supplemental instruction. However, no experimental manipulation was attempted such that the effects of increased duration could be isolated. Still, increasing the duration of intensive intervention may be one way to boost performance of this group of students who are at high risk for poor long-term reading outcomes. Finally, asking teachers who are the most experienced and knowledgeable about reading to provide instruction in Tier 3 may provide the instructional intensity that is necessary to help very poor readers make significant gains. Although the graduate students implemented the instruction with high fidelity, results of the study would likely have been different had expert teachers (Pressley et al., 2001) implemented instruction in Tier 3 and probably in Tier 2 as well.

Another unexplored option for helping struggling readers catch up is to provide the poorest readers with Tier 3 instruction immediately without delaying greater
intensity by an initial round of Tier 2 tutoring (see Vaughn et al., 2010). For a subgroup of at-risk students, Tier 2 instruction may be little more than another dose of reading failure before their eventual placement in Tier 3.

Compton et al. (2012) demonstrated that Tier 2 response data may not be necessary for accurately predicting students who are at considerable risk of being unresponsive to Tier 2 instruction. This suggests that students who require Tier 3 intervention can be identified without first participating in, and failing to respond to, Tier 2 intervention. Further research is needed to determine the adequacy of expediting the path by which certain students enter Tier 3 and what type of instruction should be provided at that more intensive tier. For example, in the present study, there was no difference between Tier 2 and Tier 3 instruction in outcomes for our most severely struggling readers. One explanation for the lack of efficacy for our Tier 3 intervention is that as part of the Tier 3 intervention, students assigned to Tier 3 repeated previously covered material, which gave that group less coverage of new material compared with students in Tier 2. Even so, we take these results to support our evolving hypothesis that Tier 3 must offer different instruction not just simply more instruction.

In considering these findings, we recognize two limitations. First, we did not attempt to enhance Tier 1 instruction. All schools in the district were required to use Macmillan/McGraw-Hill’s basal reading series Treasures (Bear et al., 2008) and were required to allocate a daily 90-minute block for reading. Because we did not provide or monitor the quality of Tier 1 instruction, we cannot ensure that the instruction at Tier 1 was strong or evidence based. Additional research is needed to experimentally evaluate the instruction at all tiers of the RTI model. Second, for reasons outlined earlier, we employed the help of trained graduate students to implement standardized instruction. Among other aspects of RTI that need further investigation, researchers and schools alike would benefit from a better understanding of who can and should be delivering upper tiers of instruction and to what extent instruction can and should be individualized for all at-risk students.

NOTES
This research was supported in part by Grant R324G060036 from the Institute of Education Sciences, U.S. Department of Education, and Core Grant HD15052 from the National Institute of Child Health and Human Development to Vanderbilt University. Statements do not reflect the position or policy of these agencies, and no official endorsement by them should be inferred.

REFERENCES


Submitted April 6, 2012
Final revision received September 27, 2012
Accepted October 8, 2012

JENNIFER K. GILBERT (corresponding author) is a research associate in the Department of Special Education at Vanderbilt University, Nashville, TN, USA; e-mail jennifer.k.gilbert@vanderbilt.edu.

DONALD L. COMPTON is a professor in the Department of Special Education at Vanderbilt University, Nashville, TN, USA; e-mail donald.l.compton@vanderbilt.edu.

DOUGLAS FUCHS and LYNN S. FUCHS are professors and the Nicholas Hobbs Chairs of Special Education and Human Development at Vanderbilt University, Nashville, TN; email douglas.fuchs@vanderbilt.edu and lynn.fuchs@vanderbilt.edu.

BOBETTE BOUTON is a doctoral student at the University of Georgia, Athens, USA; e-mail bdbouton@uga.edu.

LAURA A. BARQUERO and EUNSOO CHO are doctoral students at Vanderbilt University, Nashville, TN, USA; e-mail laura.a.barquero@vanderbilt.edu and eunsoo.cho@vanderbilt.edu.