IDENTIFYING READING DISABILITIES BY RESPONSIVENESS-TO-INSTRUCTION: SPECIFYING MEASURES AND CRITERIA

Douglas Fuchs, Lynn S. Fuchs, and Donald L. Compton

Abstract. First, we describe two types of assessment (problem solving and standard treatment protocol) within a "responsiveness-to-instruction" framework to identify learning disabilities. We then specify two necessary components (measures and classification criteria) to assess responsiveness-to-instruction, and present pertinent findings from two related studies. These studies involve databases at grades 1 and 2, which were analyzed to compare the soundness of alternative methods of assessing instructional responsiveness to identify reading disabilities. Finally, conclusions are drawn and future research is outlined to prospectively and longitudinally explore classification issues that emerged from our analyses.

DOUGLAS FUCHS, Ph.D., is professor, Peabody College of Vanderbilt University.
LYNN S. FUCHS, Ph.D., is professor, Peabody College of Vanderbilt University.
DONALD L. COMPTON, Ph.D., is assistant professor, Peabody College of Vanderbilt University.

Over the 25-year history of the Individuals with Disabilities Education Act (IDEA), the number of students identified as having learning disabilities (LD) has increased dramatically. Prior to 1970, students with LD were rarely identified. Now, they comprise more than 50% of all children with disabilities, or 5% of the school population (U.S. Department of Education, 2000). The dramatic increase in the prevalence of LD has raised concerns about the methods by which these children are identified.

This concern, we believe, is well founded. Because LD is defined as unexpected failure to learn, the discrepancy between intelligence and achievement has been the keystone in the process by which LD is typically identified. Yet, the measurement of discrepancy is problematic because of the poor reliability of difference scores (Reynolds, 1984), and because practitioners' use of varying discrepancy formulae and test instruments tend to identify different students (e.g., Shepard, Smith, & Vojir, 1983). Moreover, research documents similar underlying deficits in children with reading difficulties whether or not they demonstrate discrepancies between intelligence and achievement (Fletcher et al., 1998; Fletcher et al., 1994; Francis, Fletcher, Shaywitz, Shaywitz, & Rourke, 1996; Velutino et al., 1996).

These and other problems have prompted calls for alternative identification methods (e.g., Lyon et al., 2001; Siegel, 1989). One alternative approach is responsiveness-to-instruction, or RTI. With RTI, students are identified as LD when their response to generally effective instruction (i.e., instruction to which most children respond) is dramatically inferior to that of their peers. The basic assumption is that RTI can differentiate between two explanations of low achievement: poor instruction versus disability. If a child is nonresponsive to instruction that benefits a majority of stu-

dents, the assessment process eliminates poor instruction as an explanation for the child's inadequate growth. Instead, it suggests that disability is responsible and that specialized intervention is necessary to boost academic achievement and chances for post-school success.

RTI has generated considerable attention. The U.S. Department of Education's Office of Special Education Programs recently sponsored a series of white papers and an LD Summit (see Bradley, Danielson, & Hallahan, 2002), partly to explore the viability of RTI. The President's Commission on Excellence in Special Education (2002) and a National Academy of Sciences committee on overrepresentation of minority students in special education (Donovan & Cross, 2002) also encouraged consideration of its use. Moreover, an entire issue of *Learning Disabilities Research and Practice* (Vaughn & Fuchs, 2003) was recently devoted to the topic.

Despite this mostly positive attention, many questions about RTI remain unanswered. For example, the social consequences of such a reorientation to LD identification, including prevalence rates, equity issues, and prevention outcomes, are yet to be studied. There are questions, too, about what measures of and criteria for instructional responsiveness should be used to yield reliable and valid decision-making.

In this article, we focus on assessment for identification of reading disability. By some estimates (Lyon, l995), 80% of students with LD suffer their most serious academic difficulties in reading. Although, in the earliest grades, this mostly involves word analysis and word identification, eventual problems include reading fluency and comprehension (Gough, 1996; Perfetti, Marron, & Foltz, 1996; Shankweiler et al., 1999), which grow more serious as the school curriculum focuses increasingly on reading for meaning and for learning new information in the later grades.

We begin by explaining conceptual and technical strengths and weaknesses of two forms of RTI for reading disability identification. We then specify the components necessary to assess instructional responsiveness, and present data from two recent and pertinent studies, in which we explore the technical soundness of alternative operationalizations of instructional responsiveness. We conclude by outlining prospective and longitudinal research to examine identification and classification issues.

READING DISABILITY AS RTI: TWO CONCEPTUAL APPROACHES

Problem-Solving in General Education

An RTI approach to identifying disability is rooted in a 1982 National Research Council study (Heller, Holtzman, & Messick), which proposed that the validity of any special education classification must be judged according to three criteria: (a) that mainstream education was generally effective; (b) that special education improved student outcomes, thus justifying the classification; and (c) that the assessment process used for identification was valid. Only when all three criteria are met, claimed Heller et al., was a special education classification justifiable.

Fuchs (1995) borrowed the Heller et al. (1982) framework (see also Fuchs & Fuchs, 1998) to specify a three-phase process to assess disability. In Phase I, the rate of growth of all students in a mainstream classroom is tracked. The purpose of such classwide assessment is to determine whether the instructional environment is sufficiently nurturing to expect student progress. If, across all students, the mean rate of growth is low in comparison to other classes of children in the same building, the same district, or the entire nation, the appropriate decision would be to intervene at the classroom level to develop a stronger instructional program for all.

After establishing that classroom instruction is generally effective, Phase II assessment commences with the identification of students whose level of performance and rate of improvement are well below those of classroom peers. The purpose of this assessment, therefore, is to identify a subset of children whose potential academic failure is signaled by their unresponsiveness to generally effective instruction. For only these children, the next phase, Phase III assessment, includes problem-solving and systematic tryouts of individualized adaptations in the mainstream setting. The purpose of problem solving and adaptations is to determine whether the general education classroom can be transformed into a productive learning environment for these at-risk students. Only when such adaptations fail to improve student growth do practitioners consider special services. The assumption is that if the individualized adaptations do not produce growth for the at-risk students, some inherent deficit or disability is probably making it difficult for them to benefit.

To conduct Phase I, II, and III assessments, Fuchs (1995) suggested curriculum-based measurement (CBM; Deno, 1985), an approach that permits modeling of student responsiveness to instruction. In Phase I, CBM quantifies "classroom instructional quality" as mean performance level and growth rate for the entire class. In Phase II, "risk" is defined as a dual discrepancy (on CBM performance level and CBM growth rate) between the targeted at-risk student and classmates. In Phase III, CBM is used to index "responsiveness to classroom adaptations," with the goal of boosting the at-risk student's CBM level and rate within the range of the

class mean. Fuchs provided data to show how CBM meets important standards with respect to Heller et al.'s (1982) third criterion: that the assessment process used for classification, requiring judgments about the quality of the instructional setting and the student's responsiveness in that setting, is accurate and meaningful.

Standard Treatment Protocol

To address at-risk students' learning problem in general education, Fuchs (1995) proposed a series of adaptations teachers might incorporate in a routine way. More recently, others have reformulated Phase III in Fuchs's model to more strongly emphasize remediation of at-risk students' difficulties. Sometimes this is attempted through an iterative problem-solving process (e.g., Grimes, 2000; Marston et al., 2003). More commonly, an intensive fixed-duration trial (e.g., 10-15 weeks) of small-group or individual tutoring is used, involving a validated standard treatment protocol (e.g., Al Otaiba & Fuchs, 2004; McMaster, Fuchs, Fuchs, & Compton, in press; Vellutino et al., 1996). If the student responds to an intensive treatment trial, she is seen as remediated and disability-free and is returned to the general education classroom for instruction. If, on the other hand, she is non-responsive, a disability is suspected and further evaluation is warranted.

A recent study by Vaughn, Linan-Thompson, and Hickman-Davis (2002) illustrates this more recent standard treatment protocol approach to RTI. Second-grade students at-risk for reading disability were assessed and provided 10 weeks of supplemental, small-group reading instruction. Afterwards, all who met a priori cut-points were no longer included in the supplemental instruction; remaining students were regrouped and provided another 10 weeks of instruction. This continued for 30 weeks, when the subset of students who still had not met criteria for dismissal from supplemental instruction (25% of the original sample) were considered for special education.

This relatively intensive three-phase approach transforms an identification process into prevention. Variations by others on this preventive approach include different numbers of tiers, or phases, and different types of activities occurring at the various tiers (see Fuchs, Mock, Morgan, & Young, 2003, for discussion).

Conceptual and Technical Distinctions

Problem solving in general education and use of standard treatment protocols represent two approaches to RTI. They differ both conceptually and with respect to technical issues. Each, for example, has its own implicit meaning of "responsiveness/non-responsiveness." Use of a standard treatment protocol provides a very rigorous test for non-responders and the presence of disability. Students, like those in Vaughn et al.'s (2002) study, participate in a research-backed, intensive, and iterative instructional process. In such circumstances, it makes little sense to point to poor or inadequate instruction as a cause of non-responsiveness. It makes more sense to consider disability as a cause. At the same time, use of a standard treatment protocol raises the question: Is it possible that some children who are responsive to instruction in a second or third tier of a multi-tier approach *still* have disabilities and, once returned to general education instruction without the intensity and systematicity of the standard treatment protocol, again demonstrate the same learning problems that first marked them as candidates for participation in the standard treatment protocol? In short, whereas the standard treatment protocol approach is likely to identify "true" non-responders, is it also likely to identify "false" negatives? For example, in the Vaughn et al. study, a subset of children who met criteria for dismissal from intensive tutoring subsequently failed to thrive in general education and eventually required additional attention.

By contrast, an at-risk student's responsiveness to general education with individualized adaptations suggests that adequate learning will continue without further intervention. Students in a generally effective instructional classroom with adaptations, whose learning is much worse than that of classroom peers, are likely to require the intensity of instruction special education is meant to provide. Moreover, defining "intervention" and "responsiveness/non-responsiveness" in general education presumes that disability should be assessed as it occurs under "normal" conditions: in the mainstream setting. This parallels contexts in which other psychological conditions are diagnosed. Ruling out disability only after intensive effort improves a condition seems akin to concluding that a patient never had cancer because surgery restored her to health.

Regarding technical issues, problem solving and standard treatment protocol approaches create different challenges. Relying on general education to assess responsiveness to instruction has the advantage of a normative framework referenced to the typical population. That is, responsiveness to generally effective instruction can be estimated for all students so that a normative profile can be generated to describe the full range of response. With general education instruction as the intervention, traditional cut-points (e.g., 1.5 standard deviations below the mean) may be used to define disability. Such an approach requires measurement of all students. By contrast, it seems unlikely that a normative framework may be applied to the standard treatment protocol approach. Thus, logistics and logic seem to argue against exposing the full range of stu-

dents to an intensive tutoring regimen for the purpose of producing a normative profile. In all likelihood, practitioners would need to rely on a normative framework restricted to very poor readers, a proposition requiring empirical validation.

In comparison to the standard treatment protocol approach, problem solving is usually associated with a lower bar to determine non-responsiveness and easier access to special education. Assuming that special education is effective, this helps ensure that all children with special needs receive appropriate services. Yet, relatively easy access to special education can, in some cases, reflect a "rush to judgment" and identification of "false positives," or children who are incorrectly identified and labeled. The standard treatment protocol approach, by contrast, tends to provide more intensive instruction, to which many children respond positively. However, it is also more likely to produce "false negatives," or students with disabilities who improve during intensive tutoring only to be returned to general education where they fail once again. In selecting between these two approaches, it may be necessary to determine whether one's primary intent is identification or prevention.

READING DISABILITY AS RTI: TWO ASSESSMENT COMPONENTS

Regardless of which RTI approach is adopted, two components of the assessment process must be specified. First, methods must be determined for measuring students' response to instruction. That is, measures must be specified for tracking responsiveness, and so must the frequency with which the measures are administered. Second, once student responsiveness has been quantified, a criterion must be applied for defining non-responsiveness. Below such a criterion, students are identified as having reading disabilities.

Prior Research on Measuring and Defining Non-Responsiveness

Various methods are available for specifying these two assessment components. Vellutino et al. (1996) tested students on subtests of the Woodcock Reading Mastery Tests several times over the course of a multiyear study. To establish a cut-point for responsiveness, they rank-ordered slopes representing children's growth in responsiveness to tutoring, performed a "median split" on the slopes, and designated the bottom half as non-responsiveness. Similarly, Torgesen and colleagues (2001) evaluated student performance at the end of treatment on the subtests of the Woodcock Reading Mastery Tests, designating non-responsiveness as failing to achieve "normalized" status; that is, a word-reading standard score of 90 or better. Finally, Good, Simmons, and Kame'enui (2001), like Torgesen et al., also specified non-responsiveness in terms of posttreatment status. However, their approach involves a criterion-referenced "benchmark" associated with future reading success.

Speece and Case (2001) took yet a different tack. They adopted frequent measurement using CBM so that non-responsiveness could be identified earlier in the school year than was possible with the Vellutino et al., Torgesen et al., or Good et al. methods. Speece and Case applied a "dual discrepancy" criterion. Non-responders were students whose slope and level of performance fell at least 1 standard deviation below their class mean. This dual-discrepancy approach could also be determined with respect to school, district, or national norms or using benchmark cut-points associated with future school success.

Many other options exist for measuring and defining students' non-responsiveness to instruction. Unfortunately, few studies have explored these alternatives.

Our Research on Measuring and Defining Non-Responsiveness

To provide information about how to identify responders and non-responders, we retrospectively analyzed the data from two reading intervention studies – both designed in parallel fashion, involving a standard treatment protocol. One study was conducted in first grade; the other in second grade. In each grade, we identified students for intensive tutoring in a rather unique manner. Instead of assessing and identifying them in the beginning of the school year, 20 first- and second-grade teachers implemented Peer-Assisted Learning Strategies (PALS; Fuchs & Fuchs, in press; Fuchs, Fuchs, Mathes, & Simmons, l997; Fuchs et al., 2001), a validated classroom-based reading program. In each of the 20 classes, we designated a subset of children as at risk based on beginning-of-school-year screenings: approximately 40% of the full sample of first-graders (the lowest eight students per class on letter naming fluency) and 30% of second-graders (the lowest six students per class on CBM). We monitored these at-risk students' responsiveness to the PALS program. We also monitored the responsiveness of typically achieving children. At the end of the first semester, we identified non-responsive students whose performance was substantially below that of classroom peers.

To monitor progress at grade 1, we collected weekly data in two areas: word identification and word attack. To index word identification, we measured students on alternate forms of the Dolch word list, where students had 1 minute to read high-frequency words. To track the development of word attack skills, we measured

students on alternate forms of the nonsense word fluency measure (see the Dynamic Indicators of Basic Early Literacy Skills; DIBELS; Good et al., 2001). With nonsense word fluency, students are given lists of consonant-vowel-consonant pseudo-words and are instructed to say sounds or decode the pseudo-word. The score is the number of sounds read correctly (with three sounds awarded for a correctly decoded pseudo-word) in 1 minute. Second-graders' reading development was monitored with CBM oral reading fluency (Deno, 1985).

Using these progress-monitoring data, we calculated dual discrepancies relative to classroom peers and the entire experimental sample. Students who were non-responsive to PALS were at least .5 standard deviations below the reference groups on both measures in grade 1 and on CBM in grade 2. Using this method, we identified 54 first-graders and 64 second-graders requiring additional attention. This represented about 13% and 10% of the experimental groups in first and second grade, respectively. These children were then assigned randomly to intensive tutoring or to continue in PALS. The subset of students assigned to intensive tutoring, 36 of the 54 first-graders and 48 of the 64 second-graders, are the children on whom we conducted the analyses described below.

At both grade levels, the tutoring activities addressed phonological awareness, letter-sound recognition, decoding, sight-word recognition, fluency building, and sentence and story reading. Tutoring was conducted for 10-12 weeks, 30-35 minutes per session. At grade 1, the one-to-one sessions were conducted three times a week. At grade 2, students were assigned randomly to small-group instruction or individual tutoring, which, in either case, was conducted four times a week. Throughout the tutoring, the weekly progress monitoring continued.

Below, we describe additional study procedures and summarize the findings separately for the grade 1 and grade 2 databases. These analyses were conducted retrospectively. Therefore, our methods for judging instructional responsiveness, and our strategies for assessing the validity of the methods, were limited to variables in the database. The reader should be mindful that these analyses address responsiveness to an intensive standard treatment protocol conducted during the second semester, not to the implementation of PALS in the general education classroom during the first semester.

First-Grade Study Procedures and Findings

Study procedures. At grade 1, responsiveness to a standard treatment protocol was judged using four methods. The first two were modeled after Vellutino et al. (l996), using median splits on slopes calculated over the course of the tutoring: one on the Dolch weekly monitoring data; the other on the nonsense word fluency weekly monitoring data. The remaining two methods were based on students' posttreatment status. Using Torgesen et al.'s (2001) framework, one criterion for determining responsiveness was achieving "normalized" posttreatment status; that is, a standard score of 90 or greater on the word reading score of the Woodcock Reading Mastery Tests. The other posttreatment status criterion was based on the DIBELS's year-end first-grade benchmark of 40 words read correct from text in 1 minute (Good et al., 2001). We refer to these four methods of assessing responsiveness, respectively, as (a) Dolch slope median split, (b) nonsense word fluency slope median split, (c) normalized posttreatment status, and (d) benchmark posttreatment status.

To explore the validity of these methods, we created responsive and non-responsive groups using each method. Then, for each method, we contrasted the outcome (May) performance and amount of growth (May raw score minus September raw score) of the responsive and non-responsive groups on the various reading measures in our extant database. Our assumption was that the more valid and preferred methods for judging instructional responsiveness would better differentiate the outcomes and growth of the responsive and non-responsive groups. For the May outcome performance, we examined students' (a) standard scores on the Woodcock Reading Mastery Tests (Word Identification and Word Attack), (b) spelling standard scores on the Wechsler Individual Achievement Test, and (c) fluency and (d) comprehension raw scores on the Comprehensive Reading Assessment Battery. The Comprehensive Reading Assessment Battery requires students to read two 400-word passages aloud. After reading each passage, students answer 10 short-answer questions that address idea units of high thematic importance.

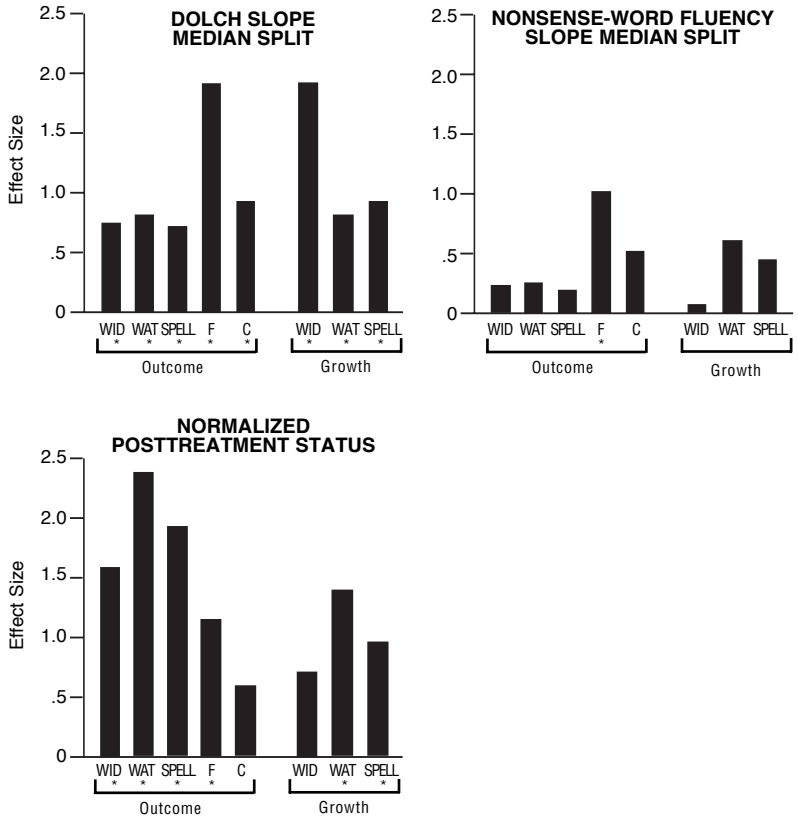
Findings. The proportion of tutored children designated non-responsive was 47.2 for the Dolch slope median split, 47.2 for the nonsense word fluency slope median split, 16.7 for the normalized posttreatment status, and 100 for the benchmark posttreatment status. By design, the median split methods identified approximately half the tutored sample, which translates into 3.5% of the full experimental sample. The two posttreatment status methods resulted in dramatically different prevalence rates of non-responders: 1.4% of the full experimental sample for normalized posttreatment status vs. 8.4% for benchmark posttreatment status. Normalized posttreatment status proved the most lenient criterion (i.e., lowest proportion of non-responders), whereas the benchmark posttreatment criterion was the most stringent criterion (i.e., highest

proportion of non-responders). Effect sizes and statistical significance (represented by asterisks) are shown in Figure 1.

In terms of how well the alternative methods differentiated responders' and non-responders' outcomes and growth, the two slope criteria performed differently (see Figure 1). Dolch slope median split fared relatively well, identifying responsive and non-responsive groups that performed statistically significantly differently, with large effect sizes, on every (May) outcome variable and on every (September to May) growth variable. The average effect size for outcomes was 1.00 standard deviation; for growth, it was 1.19. On the comprehension outcome, the effect size was .90. By contrast, nonsense word fluency slope median split functioned poorly, distinguishing responsive and non-responsive groups on only one outcome (text reading fluency) and on none of the growth measures. The average effect size for the outcome variables was .43; for growth, it was .36. The effect size for the comprehension outcome was .54.

Consequently, it seems that first-graders' slope on sight word recognition of Dolch high-frequency words may be a more valid overall indicator of first-graders' responsiveness to an intensive standard treatment protocol than their performance on nonsense word fluency tasks, which required decoding of closed-syllable pseudo-words. Of course, findings may be specific to the

Figure 1. Effect sizes distinguishing responders from non-responders by classification criteria and measures in grade 1. Outcome measures are the Woodcock Reading Mastery Test – Word Identification (WID) and Word Attack (WAT); Wechsler Individual Achievement Test – Spelling (SPELL); and Comprehensive Reading Assessment Battery – Fluency (F) and Comprehension (C). Growth measures are the same minus the Comprehensive Reading Assessment Battery.



measures we used for monitoring responsiveness. Some work (Morgan & Young, 2002) tentatively suggests technical problems for nonsense word fluency slope, with the relation between it and other indicators of decoding competence *decreasing* over the course of treatment. Future studies should continue to explore the technical properties of nonsense word fluency slope.

In terms of posttreament status, the normalized posttreatment status criterion fared better than the benchmark posttreatment status, as indicated in Figure 1. Judging responsiveness in terms of whether students achieved a standard score of 90 or better discriminated responsive students from non-responsive students on four of five outcomes (all but comprehension), and on two of three growth scores (word attack and spelling, but not word identification). Effect sizes were large, with averages of 1.59 for outcome and 1.05 for growth. By contrast, use of the DIBELS's benchmark criterion of 40 words read correctly from text in 1 minute (Good et al., 2001) resulted in no student being judged responsive. Hence, no data are presented for the benchmark criterion in Figure 1. While in principle, it is possible that the tutoring treatment was ineffective, this possibility is weakened by the competing responsiveness assessment methods. It is more likely that the DIBELS benchmark criterion was too stringent to discriminate responders from non-responders, at least when assessing responsiveness to an intensive standard treatment protocol for an initially very low-performing sample.

As mentioned, this database and retrospective series of analyses were limited to the variables selected for our studies. Investigators planning to prospectively explore the validity of alternative methods of judging treatment responsiveness at first grade would be well advised to include CBM's oral reading fluency in the second semester to monitor progress and to judge responsiveness. It is unfortunate that the available database cannot be used to examine the utility of CBM slope.

In summarizing, it seems useful to compare the better of the two methods for judging responsiveness based on slope (i.e., Dolch) to the better of the two methods for judging responsiveness based on posttreatment status (i.e., normalized posttreatment status). In this comparison, Dolch slope median split fared better than normalized posttreatment status in terms of the consistency with which it differentiated the performance of responsive students from that of non-responsive students. Using the Dolch approach, effects were statistically significant on every measure. Normalized posttreatment status, by contrast, failed to reliably discriminate end-of-year comprehension performance and word identification growth. Effect sizes were greater for normalized posttreatment status than for Dolch slope on outcome, but not on growth, variables. These two methods of judging responsiveness appear valid and might be used in a coordinated fashion in first grade. Future research should examine this possibility.

Second-Grade Study Procedures and Findings

Study procedures. In the grade 2 database, responsiveness to an intensive standard treatment protocol was judged in six ways. The first two methods were modeled after Vellutino et al.'s (1996) median split: one on the Woodcock word-reading gain scores; the other on CBM slope. The next two methods were based on posttreatment status. Using Torgesen et al.'s (2001) framework, one of these methods was "normalized" posttreatment status, indicated by a standard score of 90 or better on the word-reading score of the Woodcock Reading Mastery Tests. The second posttreatment status method relied on a CBM year-end grade 2 benchmark of at least 75 words read correctly from text in 1 minute. Our final two methods were also based on CBM performance: a normative criterion for expected CBM slope at grade 2 (i.e., 1.5 words' increase per week) and a combination of this CBM slope criterion and the benchmark CBM performance of 75 words correct at the end of treatment. As specified by Fuchs (1995), this last dual-discrepancy criterion designated students as non-responsive only if they failed to meet both criteria. In other words, if either growth rate or performance level was adequate, students were deemed responsive. We refer to these six methods for judging responsiveness, respectively, as Woodcock word reading gain median split, CBM slope median split, normalized posttreatment status, benchmark posttreatment status, normative CBM slope, and dual discrepancy.

The following reading outcomes were available to examine differences between responsive and non-responsive groups at grade 2. For May outcomes, the database included Word Identification and Word Attack standard scores on the Woodcock Reading Mastery Tests, spelling standard scores on the Wechsler Individual Achievement Test, and fluency and comprehension raw scores on the Comprehensive Reading Assessment Battery. For September-to-May growth (calculated as raw score gain), we used Word Identification and Word Attack scores for the Woodcock Reading Mastery Tests, spelling performance on the Wechsler Individual Achievement Test, and fluency and comprehension scores on the Comprehensive Reading Assessment Battery.

Findings. The proportion of second-graders designated as non-responders was 43.7 for word reading gain median split, 50.0 for CBM slope median split, 45.8 for normalized posttreatment status, 91.7 for

benchmark posttreatment status, and 29.2 for normative CBM slope and 29.2 for dual discrepancy. Normative CBM slope and dual discrepancy identified the same pool of students due to the stringency of the CBM benchmark posttreatment status criterion. Nevertheless, across the remaining classification methods, different proportions of students were identified as non-responsive. For example, the median split methods, by design, identified approximately half the sample (or 3.5% to 3.8% of the full experimental group), whereas the two posttreatment status methods resulted in different prevalence rates: 3.5% of the entire sample for normalized posttreatment status versus 7.0% for CBM benchmark posttreatment status. As with our first-grade study, therefore, the CBM benchmark posttreatment criterion represented a much more stringent criterion. The normative CBM slope and dual discrepancy identified the fewest students as non-responsive (1.4% of the full experimental group). This finding suggests that these initially very low-performing second-grade students grew more during tutoring than their final status might suggest. It also questions the validity of basing responsiveness criteria exclusively on posttreatment status. Thus, five methods are displayed.

In Figure 2, we present effect sizes and statistical significance (represented by asterisks) on the reading outcome and growth variables for the responder/non-responder groups as a function of classification method. The data for the normative CBM slope and dual-discrepancy methods are provided together because, as mentioned, the two methods identified identical groups of children.

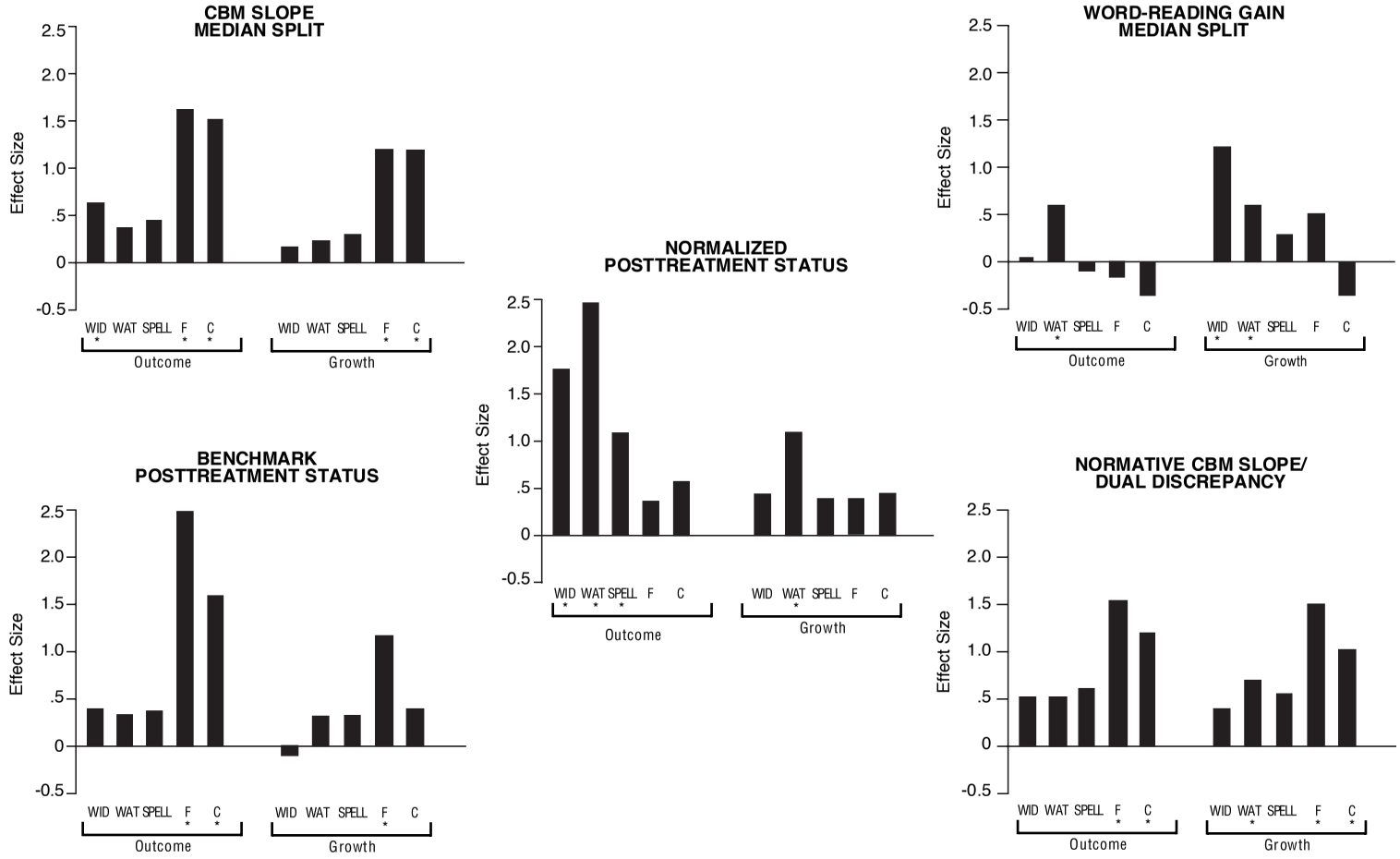
As illustrated, the CBM slope median split produced stronger differentiation between responsive and non-responsive groups than the word-reading gain median split. The responsive and non-responsive groups formed by the CBM slope median split performed statistically significantly differently on three of five outcome variables (word identification, fluency, and comprehension, but not on word attack or spelling) and on two of five growth variables (fluency and comprehension, but not on word identification, word attack, or spelling). The average effect sizes were large: .94 for outcome and 1.20 for growth, with impressive effect sizes of 1.53 and 1.20 on comprehension outcome and comprehension growth, respectively. By contrast, the Woodcock Word Identification gain median split resulted in differential performance on only the Word Attack outcome variable and on only the two Woodcock growth variables. Effect sizes were also very modest, with a mean of .01 for the outcome variables and .43 for the growth measures. Notably, effect sizes for the comprehension measures were in the wrong direction (-.34 for outcome and -.35 for growth).

The next two methods for designating responsive/non-responsive groups were based on posttreatment status: Torgesen et al.'s (2001) cut-point of 90 or higher on word reading and the second-grade CBM benchmark of at least 75 words read correctly from text in 1 minute. These two posttreatment status methods performed comparably well, although they differentiated responders and non-responders on different variables. Specifically, the normalized posttreatment word-reading method distinguished the two groups on word identification, word attack, and spelling outcome variables and on the word attack growth score. Mean effect sizes were 1.22 for outcome and .52 for growth, with corresponding effect sizes of .52 and .40 for comprehension.

By contrast, the CBM benchmark discriminated the groups on fluency and comprehension outcome variables as well as on the fluency growth score. Effect sizes were similar to those for normalized posttreatment status: 1.05 for outcome and .41 for growth. Although effect sizes for growth in comprehension were identical across the two posttreatment methods (.40), the comprehension outcome effect sizes were notably larger for CBM benchmark posttreatment status (1.63) than for normalized posttreatment status (.52). Whereas neither of the posttreatment status methods fared as well as the CBM slope median split, it should be noted that only 4 of the 36 students met the CBM benchmark criterion. This raises questions about the stringency of the CBM benchmark when used to identify non-responsiveness to intensive tutoring. These findings resemble those of the first-grade database.

The last classification method, also a variation of CBM, employed a dual discrepancy for unresponsiveness: growth less than 1.5 words per week and a posttreatment level of performance below the benchmark of 75 words read correctly. The CBM slope criterion produced the lowest percentage of unresponsive students: 29.2% (or 2.2% of the total experimental sample), as opposed to 43.7% for word reading gain median split (3.5% of the experimental sample), 50.0% for CBM slope median split (3.8% of the experimental sample), 45.8% for normalized posttreatment status (3.5% of the experimental sample), and 91.7% for CBM benchmark posttreatment performance (7.0% of the experimental sample). Thus, when compared to typically performing students' responsiveness to general education, many tutored students demonstrated respectable rates of improvement, suggesting an absence of disability among many of the students even though they failed to achieve posttreatment criteria for adequate performance. As the benchmark associated with a good prognosis increases with each grade, questions arise about whether these children must remain

Figure 2. Effect sizes distinguishing responders from non-responders by classification criteria and measures in grade 2. Outcome and growth measures are the Woodcock Reading Mastery Test – Word Identification (WID) and Word Attack (WAT); Wechsler Individual Achievement Test – Spelling (SPELL); and the Comprehensive Reading Assessment Battery – Fluency (F) and Comprehension (C).



in intensive tutoring and, if so, for what length of time, and what resources might pay for the service. These conceptual and policy issues should be considered carefully before an RTI framework for reading disability classification is complete.

In any case, normative CBM slope/dual discrepancy fared well in terms of the consistency and magnitude of effects in discriminating responsive from non-responsive students. This classification method produced statistically significant effects on five variables: two outcomes (fluency and comprehension) and three growth measures (word attack, fluency, and comprehension). Average effect sizes were large: .85 for outcome variables and .84 for growth. Effect sizes for comprehension outcome and growth were 1.15 and 1.05, respectively.

Two additional points are worth noting. First, as mentioned, the dual-discrepancy method resulted in groups identical to those identified based on slope alone. This was because few students achieved the posttreatment CBM benchmark of at least 75 words read correctly in 1 minute. Consequently, the dual criterion was unnecessary; normative slope served to differentiate the groups. Second, dual discrepancy fared no better than the CBM slope median split. The dual-discrepancy method, as conceptualized by Fuchs and Fuchs (l998) and studied by Speece and Case (2001), establishes criteria for slope and level relative to those of classroom peers, not with respect to the broad, normative framework used for the present analysis. Therefore, we cannot comment on the reasonableness of cut-scores framed with reference to the local context. Moreover, lower benchmark cut-points employed within a dual-discrepancy approach would have produced different groups of students from those based on a focus on only normative CBM slope. It would be interesting to determine a CBM benchmark that actually forms different groups for the two approaches and to explore how the two groups differ.

CONCLUSIONS

These findings are preliminary because of small sample sizes and the retrospective nature of the analyses. Findings require corroboration with larger samples followed prospectively and longitudinally across the primary grades to investigate long-term outcomes. For now, we tentatively draw several conclusions across our two databases.

First, alternate methods of assessing responsiveness produce different prevalence rates of reading disability and different subsets of unresponsive children. This is important because a major criticism of IQ-achievement discrepancy as a method of LD identification is the unreliability of the diagnosis. Practitioners relying on an assortment of assessment procedures in an RTI framework may produce similarly unreliable diagnoses. To develop more consistent identification procedures, researchers must explore the soundness of various methods. At the same time, however, different assessment methods demonstrate differential utility in distinguishing responsive and non-responsive groups on different components of beginning reading. For this reason, consistency in identifying non-responders across the various components of beginning reading skill is an important criterion for selecting a valid assessment approach. Among the alternatives we explored, Dolch slope median split was the clear winner in terms of its consistency in grade 1. Thus, it discriminated responsive/non-responsive groups on all five outcome variables and all three growth variables. At second grade, no approach differentiated responders from non-responders on all outcome and growth variables. However, CBM slope median split and normative CBM slope/dual discrepancy fared best with respect to consistency. Thus, CBM slope median split differentiated the two groups on three of five outcome variables and two of five growth variables. Normative CBM slope/dual discrepancy differentiated the groups on two of five outcome and three of five growth variables.

Second, CBM benchmark posttreatment status (as defined in our analyses) was a considerably more stringent criterion than the other methods. It did not produce a single responder at grade 1, and only four responders at grade 2. The question is whether the cut-points of 40 words read correctly per minute at grade 1 and 75 words read correctly per minute at grade 2 are too high to define responsiveness to intensive standard treatment protocols. The answer might depend on how students are selected to participate in intensive tutoring. In our work, children identified for tutoring had already demonstrated poor responsiveness during an entire semester of PALS, a validated classroom reading program. In others' work, children have been chosen for tutoring based on September screening scores. September screening will surely produce more false positives for risk status (Jenkins & O'Connor, 2002). With a higher proportion of false positives in the tutoring treatment, a better rate of responsiveness, and more defensible grounds for use of CBM posttreatment benchmarks, can be predicted.

A third conclusion drawn across the first- and second-grade studies concerns the use of posttreatment status as a means of indexing responsiveness. As represented by Torgesen et al.'s (2001) cut-point of a standard score of 90 or better on the Woodcock Word Identification score, normalized posttreatment status differentiated responsiveness from non-responsiveness on posttreatment outcome measures better than on

growth measures. This finding should come as no surprise given that judging responsiveness by means of posttreatment status fails to consider amount of learning. At the same time, Dolch slope (at grade 1) and normative CBM slope (at grade 2) differentiated responsive from non-responsive students' performance equally well on outcome and growth variables, suggesting the potential utility of slope as an index of responsiveness. Of course, in these analyses, outcome and growth were defined within a short timeframe. The real key is formulating optimal cut-points to identify the children who fare worst over the course of their educational experience, and for whom reading, especially reading for meaning, represents a life-long skill deficit that results in poor post-school outcomes.

Our final conclusions concern reading comprehension. In the first-grade database, Dolch median split produced the largest difference between responsive/non-responsive groups on comprehension, where only outcome (not growth) information was available. At second grade, CBM slope median split and benchmark posttreatment status yielded the largest between-group differences on comprehension outcome; CBM slope median split and normative CBM slope/dual discrepancy produced the largest between-group differences on comprehension growth. At grade 2, monitoring student responsiveness with CBM was clearly superior to Woodcock Word Identification in terms of its correspondence to reading comprehension, at least as operationalized in these studies.

Rather than regarding these conclusions as written in stone, we offer them as reasonable hypotheses with which to begin prospective, systematic, and longitudinal research on the utility of alternative assessments in an RTI framework. At least three major components of such assessments need to be examined. First, research should explore how classification varies as a function of the nature of the treatment. It is likely that the criteria by which reading disability is predicted will require different cut-points when responsiveness is assessed in general education versus in intensive tutoring. In addition, keeping the nature of treatment constant, researchers must give serious thought to how children enter responsiveness assessment. The utility of alternative approaches to assessment is likely to vary as a function of entry criteria.

The second component of future research concerns the nature of the measures used and the frequency of assessment. A third component addresses the criteria applied to define unresponsiveness. As demonstrated in the analyses of our first-grade and second-grade databases, different measurement systems using different criteria result in identification of different groups of students. The critical question is which combination of assessment components is most accurate for identifying children who will experience serious and chronic reading problems that prevent reading for meaning in the upper grades and impair their capacity to function successfully as adults. At this point, relatively little is known to answer this question when RTI is the assessment framework.

REFERENCES

Al Otaiba, S., & Fuchs, D. (2004). Who are the young children for whom best practices in reading are ineffective? An experimental and longitudinal study. Submitted fro publication.

Bradley, R., Danielson, L., & Hallahan, D. P. (2002). *Identification of learning disabilities: Research to practice*. Mahwah, NJ: Erlbaum.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52,* 219-232.

Donovan M. S., & Cross, C. T. (2002). *Minority students in special and gifted education.* Washington, DC: National Academy Press.

Fletcher, J. M., Shaywitz, S. E., Shankweiler, D. P., Katz, L., Liberman, I. Y., Stuebing, K. K., Francis, D. J., Fowler, A. E., & Shaywitz, B. A. (1994). Cognitive profiles of reading disability: Comparisons of discrepancy and low achievement definitions. *Journal of Educational Psychology, 86,* 6-23.

Fletcher, J. M., Francis, D. J., Shaywitz, S. E., Lyon, G. R., Foorman, B. R., Stuebing, K. K., & Shaywitz, B. A. (1998). Intelligent testing and the dual discrepancy model for children with learning disabilities. *Learning Disabilities Research and Practice, 13,* 186-203.

Francis, D. L., Fletcher, J. M., Shaywitz, B. A., Shaywitz, S. E., & Rourke, B. P. (1996). Defining learning and language disabilities: Conceptual and psychometric issues with the use of IQ tests. *Language, Speech, and Hearing in Schools, 27,* 132-143.

Fuchs, D., & Fuchs, L. S. (in press). Peer-assisted learning strategies: Promoting word recognition, fluency, and reading comprehension in young children. *The Journal of Special Education.*

Fuchs, D., Fuchs, L. S., Mathes, P., & Simmons. D. (1997). Peer-assisted learning strategies: Making classrooms more responsive to student diversity. *American Educational Research Journal, 34,* 174-206.

Fuchs, D., Fuchs, L. S., Yen, L., McMaster, K., Svenson, E., Yang, N., Young, C., Morgan, P., Gilbert, T., Jaspers, J., Jernigan, M., Yoon, E., & King, S. (2001). Developing first-grade reading fluency through peer mediation. *Teaching Exceptional Children, 34*(2), 90-93.

Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-Intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research and Practice, 18*(3), 157-171.

Fuchs, L. S. (1995, May). *Incorporating curriculum-based measurement into the eligibility decision-making process: A focus on treatment validity and student growth.* Paper prepared for the National Academy of Sciences Workshop on Alternatives to IQ Testing, Washington, DC.

Fuchs, L. S., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research and Practice, 13,* 204-219.

Fuchs, L. S., Fuchs, D., & Speece, D. L. (2002). Treatment validity as a unifying construct for identifying learning disabilities. *Learning Disability Quarterly, 25,* 33-45.

Good, R. H. III, Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of

fluency-based indicators of foundational reading skills for third grade high-stakes outcomes. *Scientific Studies of Reading, 5,* 257-288.

Gough, P. B. (1996). How children learn to read and why they fail. *Annals of Dyslexia, 46,* 3-20.

Grimes, J. (2002). *Responsiveness to interventions: The next step in special education identification, service and existing decision making*. Revision of a paper written for the Office of Special Education Programs, U.S. Department of Education, and presented at its LD Summit conference, Washington, DC.

Heller, K. A., Holtzman, W. H., & Messick, S. (Eds.). (1982). *Placing children in special education: A strategy for equity*. Washington, DC: National Academy Press.

Jenkins, J. R., & O'Connor, R. E. (2002). Early identification and intervention for young children with reading/learning disabilities. In R. Bradley, L. Danielson, & D. P. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 99-149). Mahwah, NJ: Erlbaum.

Lyon, G. R. (1995). Research initiatives in learning disabilities: Contributions from scientists supported by the National Institute of Child Health and Human Development. *Journal of Child Neurology, 10* (suppl. 1), S120-S126.

Lyon, G. R., Fletcher, J. M., Shaywitz, S. E., Shaywitz, B. A., Torgesen, J. K., Wood, F. B., Schulte, A., & Olsen, R. (2001). Rethinking learning disabilities. In C.E. Finn, Jr., R.A.J. Rotherham, & C.R. O'Hokanson, Jr. (Eds.), *Rethinking special education for a new century* (pp. 259-287). Washington, DC: Fordham Foundation.

Marston, D. (2001). *A functional and intervention-based assessment approach to establishing discrepancy for students with learning disabilities*. Paper written for the Office of Special Education Programs, U.S. Department of Education and presented at its LD Summit conference, Washington, DC.

McMaster, K. L., Fuchs, D., Fuchs, L. S., & Compton, D. L. (in press). Responding to non-responders: An experimental field trial of identification and intervention methods. *Exceptional Children*.

Morgan, P., & Young, C. (2002, June). *Effects of tutoring on the reading performance of treatment resistant children*. Poster presented at the annual meeting of the Society of the Scientific Study of Reading, Chicago.

Perfetti, C. A., Marron, M. A., & Foltz, P. W. (1996). Sources of comprehension failure: Theoretical perspectives and case studies. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and interventions* (pp. 137-165). Mahwah, NJ: Erlbaum.

President's Commission on Excellence in Special Education. (2002). *A new era: Revitalizing special education for children and their families*. Washington, DC: Author.

Reynolds, C. R. (1984). Critical measurement issues in learning disabilities. *Journal of Special Education, 18,* 451-476.

Shankweiler, D., Lundquist, E., Katz, L., Stuebing, K. K., Fletcher, J. M., Brady, S., Fowler, A., Dreyer, L. G., Marchione, K. E., Shaywitz, S. E., & Shaywitz, B. A. (1999). Comprehension and decoding: Patterns of associations in children with reading difficulties. *Scientific Studies of Reading, 3,* 69-94.

Shepard, L. A., Smith, M. L., & Vojir, C. P. (1983). Characteristics of pupils identified as learning disabled. *American Educational Research Journal, 20,* 309-332.

Siegel, L. S. (1989). IQ is irrelevant to the definition of learning disabilities. *Journal of Learning Disabilities, 22,* 469-478, 486.

Speece, D. L., & Case, L. (2001). Classification in context: An alternative to identifying early reading disability. *Journal of Educational Psychology, 93,* 735-749.

Torgesen, J. K., Alexander, A., Wagner, R., Rashotte, C., Voeller, K., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities, 34,* 33-58.

U.S. Department of Education. (2000). *Twenty-first annual report to Congress on the implementation of the Individuals with Disabilities Act*. Washington, DC: U.S. Government Printing Office.

Vaughn, S. R., & Fuchs, L. S. (Eds.). (2003). Redefining learning disabilities as inadequate response to treatment: The promise and potential problems. *Learning Disabilities Research and Practice, 18*(3), 137-146.

Vaughn, S., Linan-Thompson, S., & Hickman-Davis, P. (2002). Response to treatment as a means of identifying students with reading/learning disabilities. *Exceptional Children, 69,* 391-409.

Vellutino, F., Scanlon, D., Sipay, E., Small, S., Pratt, A., Chen, R., & Denckla, M. (l996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology, 88,* 601-638.

ACKNOWLEDGMENTS

This research was supported in part by Grant #H324DE000033 and Grant #324U01004 from the Office of Special Education Programs, U.S. Department of Education, and Grant HD 15052 from the National Institute of Child Health and Human Development, all to Vanderbilt University. The article does not necessarily reflect positions or policies of the funding agencies, and no official endorsement by them should be inferred.

Portions of this work were presented at a Center for Improvement of Early Reading Achievement (CIERA) conference on Assessment of Reading Comprehension, Ypsilanti, MI, 2002, and at the Pacific Coast Research Conference, Coronado, CA, 2004.

Requests for reprints should be addressed to: Douglas Fuchs, Vanderbilt University, Peabody #328, 230 Appleton Place, Nashville, TN 37203-5721; or doug.fuchs@vanderbilt.edu.