Oral Reading Fluency Assessment: Issues of Construct, Criterion, and Consequential Validity

Sheila W. Valencia
University of Washington, Seattle, Washington, USA

Antony T. Smith
University of Washington, Bothell, Washington, USA

Anne M. Reece
University of Washington, Seattle, Washington, USA

Min Li
University of Washington, Seattle, Washington, USA

Karen K. Wixson
University of Michigan, Ann Arbor, Michigan, USA

Heather Newman
University of Washington, Seattle, Washington, USA

ABSTRACT

This study investigated multiple models for assessing oral reading fluency, including 1-minute oral reading measures that produce scores reported as words correct per minute (wcpm). The authors compared a measure of wcpm with measures of the individual and combined indicators of oral reading fluency (rate, accuracy, prosody, and comprehension) to examine construct, criterion, and consequential validity. Oral reading data and standardized comprehension test scores were analyzed for students in grades 2, 4, and 6. The results indicate that assessments designed to include multiple indicators of oral reading fluency provided a finer-grained understanding of oral reading fluency and fluency assessment and a stronger predictor of general comprehension. Comparisons across grade levels also revealed developmental differences in the relation between oral reading fluency and comprehension, and in the relative contributions of oral fluency indicators to comprehension. When commonly used benchmarks were applied to wcpm scores to identify students at risk of reading difficulty, both false positives and false negatives were found. This study raises issues regarding the alignment of oral reading fluency definitions and assessment. It also raises concerns about the widespread use of wcpm measures and benchmarks to identify students at risk of reading difficulty and to plan instruction.

The past 10 years have seen an unprecedented involvement of national and state stakeholders in shaping reading instruction and accountability. In response to the National Reading Panel (NRP) report (National Institute of Child Health and Human Development [NICHD], 2000), No Child Left Behind (NCLB) legislation, Reading First, Response to Intervention, and a host of state and local policies, educators have been urged to focus on "proven practices" in the areas of phonemic awareness, phonics, comprehension, vocabulary, and fluency. They are expected to provide excellent instruction in these areas, assess student

performance in each, and use assessment results to inform instruction and evaluate student achievement.

In an effort to address assessment requirements, many schools, districts, and states have turned to a simple measure to assess reading performance. With a long history of research and use in special education, this measure involves having students read aloud from brief passages drawn from materials used in the classroom, provided in commercial testing kits, or developed independently (Deno, 2003; Fuchs & Deno, 1994; Fuchs, Fuchs, Hosp, & Jenkins, 2001; Marston, 1989; Shinn, Tindal, & Stein, 1988). The student reads for 1 minute as the teacher records errors, which produces a score reported as words correct per minute (wcpm). Over time, results of this approach to measuring oral reading have been used for a variety of purposes including screening to identify students academically at risk, placement in remedial and special education programs, monitoring student progress, improving instructional programs, and predicting performance on high-stakes assessments (Deno, 2003; Fuchs, Deno, & Mirkin, 1984; Good, Kaminski, & Dill, 2002; Klein & Jimerson, 2005; Marston & Magnusson, 1988; McGlinchey & Hixson, 2004; Shapiro, 2000). Most recently, measures using wcpm have also been associated with the assessment of oral reading fluency, one of the core strands identified in the NRP report. Although definitions of oral reading fluency vary, it generally has been defined as the ability to read text quickly, accurately, with proper phrasing and expression, thereby reflecting the ability to simultaneously decode and comprehend (Dowhower, 1987; Fuchs et al., 2001; Kuhn & Stahl, 2003; NICHD, 2000; Pikulski & Chard, 2005; Pinnell et al., 1995; Rasinski & Hoffman, 2003; Samuels, 2006; Wolf & Katzir-Cohen, 2001; Young & Bowers, 1995). Yet, assessment of oral reading fluency largely has been confined to measures of wcpm in both accountability and instructional settings (AIMSweb Progress Monitoring and RTI System, n.d.; Armbruster, Lehr, & Osborn, 2001; Hasbrouck & Tindal, 2006; Manzo, 2005, 2007; Valencia et al., 2006).

In the last several years, this approach to measuring reading has become widespread. A review of 45 Reading First plans available on the Internet reported that 39 included Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002), 11 included Texas Primary Reading Inventory (TPRI), and 5 included Phonological Awareness Literacy Survey (PALS) as a requirement or as one of several options to be used in Reading First schools (Sharp, 2004). All of these tests include a measure labeled "oral reading fluency" that is calculated as wcpm. Additionally, many teachers across the country are encouraged to administer various forms of oral reading fluency assessments using a measure of wcpm for students in grades 1–6 according to guidelines issued at state, district, and school levels (Armbruster et al., 2001; Hasbrouck & Tindal, 2006; Manzo, 2007). By some accounts, more than 2 million students in 49 states have been given oral reading tests that produce a score in wcpm (Olson, 2007). With the visibility and high-stakes nature of this form of assessment, it should be no surprise that classroom practices and programs have set their sights on increasing students' wcpm scores, causing some scholars and practitioners to decry the emphasis on speed at the possible cost of comprehension (Pearson, 2006; Pressley, Hildren, & Shankland, 2005; Rasinski, 2006; Samuels, 2007).

The prevalence of oral reading measures that rely on a metric of wcpm and the influence of these measures on accountability, curriculum, and instruction, calls for critical analysis and further research. Therefore, the purpose of this study was to investigate the construct, criterion, and consequential validity of using wcpm to measure oral reading fluency and to make educational decisions. Specifically, we compared a measure of wcpm with measures of the individual and combined indicators of oral reading fluency (rate, accuracy, prosody, and comprehension) and examined how well it identified students at risk of reading failure.

Background

Early work on using measures of wcpm as an index of reading was part of a larger program of research associated with curriculum-based measurement (CBM). Designed to assess students in several subject areas, CBMs originally used material drawn directly from classroom curricula (and later, generic reading material) with the aim of improving educational assessment and programs for special education students. The objective was to routinely monitor students' progress and to predict performance on norm-referenced tests (Deno, 1985, 2003; Deno, Mirkin, & Chiang, 1982; Fuchs & Deno, 1994; Hintze & Christ, 2004). In the area of reading, these quick and easy procedures were intended to "produce reliable and valid indicators of student growth in reading proficiency broadly defined" (Deno & Marston, 2006, p. 180). Undergirding this assessment strategy is a well-documented theoretical and empirical base that suggests that orally translating text with speed and accuracy reflects a complex, multifaceted, coordinated performance of a reader's perceptual and basic word skills, lexical representations, and connections among ideas in the text (Fuchs et al., 2001). Automaticity of low-level processes both at the word and text level, it is argued, frees up capacity for higher-level comprehension processes (LaBerge & Samuels, 1974). Consequently a measure of wcpm is considered by some to reflect both word recognition and comprehension competence (Fuchs et al., 2001; Samuels, 2006).

CBMs were designed to meet several criteria: They were to be reliable and valid, simple and efficient to administer, easily understood by teachers, and inexpensive. Deno (1985) acknowledged that CBMs did not have the same degree of content validity with respect to the ultimate goals of reading as answering comprehension questions but that they did meet these other criteria. It is worth noting that these initial measures indexing wcpm were not originally identified as oral reading fluency assessments—they were simply called CBMs; in later work Fuchs and Deno (1991) labeled the approach as general outcome measurement.

WCPM as an Indicator of Oral Reading Fluency

Although the term *reading fluency* has evolved over time (NICHD, 2000), studies that use oral reading CBMs or that employ wcpm have frequently used the metric synonymously with oral reading fluency; conversely, studies of oral reading fluency have tended to use measures of wcpm for assessment (Fuchs et al., 2001; Good, Simmons, & Kame'enui, 2001; Hasbrouck & Tindal, 2006). More recently, however, some researchers have raised questions about how well the metric represents oral reading fluency (Deno & Marston, 2006; Fuchs et al., 2001, Samuels, 2006; Valencia et al., 2006). The disjuncture seems to rest between conceptual definitions and measurement strategies. Most definitions of oral reading fluency suggest that readers must read quickly and accurately, with attention to proper phrasing and expression, and a central focus on comprehension (Dowhower, 1987; Fuchs et al., 2001; Kuhn & Stahl, 2003; NICHD, 2000; Pikulski & Chard, 2005; Pinnell et al., 1995; Rasinski & Padak, 2005; Wolf & Katzir-Cohen, 2001; Young & Bowers, 1995).

As Samuels (2007) noted, fluent readers orchestrate many skills so they can engage in word identification and comprehension simultaneously; speed, accuracy, and prosody are indicators of this simultaneity. Nevertheless, measures of wcpm take into account only two of these indicators—accuracy and rate of reading—which are fairly straightforward to measure. Rate and accuracy are combined to yield a metric of wcpm; expression and phrasing are not directly assessed. Furthermore, comprehension is not assessed on the passages actually read nor is it identified as a goal for students during the assessment process. In fact, Put Reading First (Armbruster et al., 2001), the summary publication of the NRP targeted for teachers, suggested assessing oral reading fluency using a procedure similar to CBM that simply yields a score of wcpm. It also recommends comparing students' scores with published norms or standards as do many basal reading programs and other measures of wcpm (Crawford, Tindal, & Stieber, 2001; Good et al., 2002; Good et al., 2001; Hasbrouck & Tindal, 2006).

In addition to the widely used metric of wcpm, other researchers have investigated the role of expression and phrasing, or prosody, as an integral component of oral reading fluency. Summarizing the research on oral reading fluency, Kuhn and Stahl (2003) concluded that prosody may provide the link between fluent oral reading and comprehension. They suggested that, unlike rate and accuracy, appropriate phrasing, intonation, and stress provide a clue that the reader is comprehending. Several studies have included a variety of indices of prosody in measures of oral reading fluency with most finding it to be an important indicator of fluency and contributor to comprehension, especially after the early stages of reading acquisition (Dowhower, 1987; Klauda & Guthrie, 2008; Kuhn, 2005; Rasinski, 1990; Young & Bowers, 1995; Zutell & Rasinski, 1991). Some of the best-known work in this area comes from two special studies of the National Assessment of Educational Progress (NAEP) that measured oral reading fluency of fourth-grade students using a 4-point oral reading fluency scale (Daane, Campbell, Grigg, Goodman, & Oranje, 2005; Pinnell et al., 1995). The scale focuses on phrasing, adherence to the author's syntax, and expressiveness, "those important elements of reading that are not necessarily captured when oral reading is judged solely on reading rate or accuracy" (Pinnell et al., 1995, pp. 16–17). Therefore, NAEP's measure of oral reading fluency more closely parallels definitions of prosody than measures of accuracy and rate that comprise wcpm. Furthermore, the testing procedure used in these NAEP studies included having students respond to comprehension questions, clearly communicating to them the expectation to read for understanding.

Both NAEP reports produced findings for all three variables associated with oral reading fluency—rate (words read per minute), accuracy, and "fluency" (the 4-point prosody rubric)—and examined the relations among these variables and with comprehension. The results indicated that being fluent, as defined by the NAEP rubric, did not ensure being among the most accurate or fastest readers. Therefore, prosody was thought to be distinct from accuracy and rate. The 2005 report concluded that, "the three separate oral reading abilities—rate, accuracy, and fluency—are related to each other and all three are related to reading comprehension" (Daane et al., 2005, p. v). Similarly, Deno and Marston (2006) noted,

> Were we to define fluency as the "number of words read correctly from text in one minute," we would be missing other features of fluent reading, such as prosody (i.e., reading with expression), that are not included in the CBM of oral reading. (p. 180)

Results from the NAEP special studies of oral reading fluency are frequently cited along with findings from CBM studies to make the point about the importance of fluency in skilled reading. Still, it is important to recognize that these studies measure different indicators of oral reading fluency.

WCPM as a Screening Measure: Relation to Comprehension

As we already noted, many researchers associated with CBM and others who advocate oral reading measures indexed as wcpm suggest that such assessments are good indicators of general reading expertise (e.g., Deno, 1985; Fuchs et al., 2001; Good et al., 2001); consequently measures that produce wcpm scores have gained in popularity as a way to screen students who may be at risk of reading failure. The research in this area is based largely on studies of criterion validity, some concurrent and others predictive, correlating students' wcpm scores with their performance on standardized reading tests, state standards-based tests, and other independent measures of overall reading ability (Crawford et al., 2001; Deno, 1985; Deno et al., 1982; Fuchs, Fuchs, & Maxwell, 1988; Good et al., 2001; Hintze, Shapiro, Conte, & Basile, 1997; Jenkins & Jewell, 1993; McGlinchey & Hixson, 2004; Shinn, Good, Knutson, Tilly, & Collins, 1992; Shinn et al., 1988; Stage & Jacobsen, 2001). A review of 14 studies, most including special education and mainstream students across several grade levels, found correlations between wcpm and criterion measures of reading comprehension ranging from .63 to .90, with most clustering around .80 (Marston, 1989). Another review examining correlations within a single grade level found correlations in the range of .60–.80 (Good & Jefferson, 1998).

Recent studies have raised questions about the nature of wcpm measures and their relation to comprehension. First, several studies have found considerably lower correlations (i.e., .4–.5) between wcpm and reading comprehension, accounting for only 16%–25% of the variance in comprehension scores (Kranzler, Brownell, & Miller, 1998; Pressley et al., 2005; Valencia et al., 2006; Wiley & Deno, 2005). This has prompted discussions of how methodological issues such as the range of grade levels, reading ability, passage selection, and the measurement of constrained skills such as rate and accuracy may have influenced results of prior studies and of the criterion validity of wcpm measures (Mehrens & Clarizio, 1993; Paris, Carpenter, Paris, & Hamilton, 2005; Valencia et al., 2006). Variables that have highly deviant ranges, either highly restricted or very wide, influence the magnitude of correlations that can be obtained.

Second, some researchers have identified problems associated with the use of cut scores or benchmarks associated with wcpm such as those suggested by DIBELS (Good et al., 2001) or Hasbrouck and Tindal (2006) to identify students at risk for reading failure. Several studies have suggested that using wcpm benchmarks may misidentify a substantial percentage of students who are low achieving as measured by norm-referenced reading tests (Johnson, Jenkins, Petscher, & Catts, 2009; Pressley et al., 2005; Riedel, 2007; Samuels, 2006; Schilling, Carlisle, Scott, & Zeng, 2007; Valencia et al., 2006). Such findings have implications for the consequential validity of using wcpm measures to identify students at risk and for the nature of instructional interventions. Benchmarks that result in too many false negatives (i.e., failure to identify students at risk who are at risk) leave students without much-needed intervention; conversely, benchmarks that result in too many false positives (i.e., identifying students at risk who are not at risk) waste limited resources on students who do not need them.

Third, Valencia and colleagues (2006) have raised concerns about the results of oral reading assessments that measure reading for only 1 minute. Findings from the NAEP 2002 Special Study of Oral Reading (Daane et al., 2005) indicated that students read at a faster rate for the first minute of oral reading than across an entire 198-word passage. The report suggested that measures of rate taken over very short durations may result in reader profiles that overestimate rate. Furthermore, from an ecological perspective, the majority of reading that students do both in school and out requires considerably more sustained effort and time, and it requires comprehension. The abbreviated length of 1-minute measures and the testing format used in most wcpm measures do not orient students to read for understanding nor is comprehension adequately assessed, if it is assessed at all. It seems unlikely that meaningful comprehension, especially higher levels of comprehension, could be assessed on passages requiring only 1 minute of reading. These issues are especially important in the current context where fluency assessment is high-stakes and "test prep" may be taking the form of practicing to read faster for short periods of time at the expense of stamina and understanding (Newman, 2009; Pearson, 2006).

Finally, researchers examining the developmental nature of oral reading fluency have suggested that the relation between fluency and comprehension is likely to be stronger at the early stages when children are still acquiring decoding skills and automaticity than at later stages when these skills become more fluent and there is a sharper focus on comprehension (Fuchs et al., 2001; Hosp & Fuchs, 2005; Jenkins & Jewell, 1993; Paris et al., 2005; Pikulski, 2006; Samuels, 2006;

Schwanenflugel et al., 2006; Shinn et al., 1992; Wiley & Deno, 2005; Yovanoff, Duesbery, Alonzo, & Tindal, 2005). Some also suggest that wcpm may not be a particularly good indicator of the ability to analyze more sophisticated literature or to learn new information from complex expository texts that students encounter in the intermediate and middle school years (Fuchs et al., 2001; Kranzler, Miller, & Jordan, 1999).

This overview of research regarding the development of a wcpm metric and its application to measures of oral reading fluency and screening for reading difficulty suggest several areas for further research. These frame our study. We pose four research questions aimed at the relation between wcpm and the construct of oral reading fluency, the criterion validity of wcpm with comprehension measures, and the consequences of applying wcpm results to make educational and instructional decisions. Specifically, we ask the following:

1. How does assessment of oral reading fluency using a metric of wcpm compare with a model that includes separate measures of rate, accuracy, and prosody in predicting scores on a standardized, norm-referenced measure of comprehension?
2. What is the effect of increasing oral reading time from 1 to 3 minutes on the relation between norm-referenced comprehension scores and wcpm, rate, and accuracy?
3. What are the relative contributions of rate, accuracy, and prosody in predicting comprehension, and do these patterns change across grades?
4. What are the consequences of using a wcpm measure to identify students at risk and to inform instructional decisions?

Taken together, these questions explore issues of construct, criterion, and consequential validity of approaches to assessing oral reading fluency.

Method

Participants

This study was conducted in two Pacific Northwest school districts that had diverse student populations. Students in grades 2, 4, and 6 from four elementary schools and three middle schools participated. Across the elementary schools, approximately 55% of the participants were students of color and 43% received free or reduced-price lunch; across the middle schools, approximately 51% of the students were of color and 42% received free or reduced-price lunch. Approximately one third of students were classified as English-language learners (ELLs), which was defined as speaking a language other than English and coming from a home where that language was spoken. Students classified as beginning ELLs (i.e., scoring Level 1 on the state Language Proficiency Test, a normative test of language reading proficiency), enrolled in school in the United States for less than one year, or receiving additional special education services were not included in the sample. All participating students, including ELL students, were receiving their reading instruction in mainstream, English-only classrooms. All were expected to participate, without accommodations, in the required testing associated with NCLB, Reading First, and state and district accountability systems. A total of 279 students participated in this study: 93 from grade 2, 91 from grade 4, and 95 from grade 6.

Measures

We used both norm-referenced and researcher-developed measures. Norm-referenced tests were used to assess reading comprehension. Reading passages and comprehension questions were developed and field-tested by the research team to assess accuracy, rate, prosody, and passage comprehension, and to calculate wcpm. Each measure is described in the following section.

Norm-Referenced Reading Measure—Iowa Test of Basic Skills

The reading comprehension section of Iowa Test of Basic Skills Battery (ITBS; Hoover, Dunbar, & Frisbie, 2001) was administered to all students as the dependent measure of reading comprehension. It includes literal and inferential comprehension of passages from a range of genres and topics. Internal consistency is reported in the range of the .90s; normal curve equivalent (NCE) scores were used in the analysis.

Prosody Rubric

We adapted the fourth-grade NAEP oral reading fluency scale to reflect the reading development of students in the participating grades and used these measures to holistically assess phrasing, adherence to the author's syntax, and expression (Daane et al., 2005; Pinnell et al., 1995). As noted previously, although NAEP labeled the rubric a fluency scale, it is used to assess aspects of oral reading that more closely align with prosody. Therefore, for the purposes of this study, we refer to the variable defined by the NAEP scale as prosody to distinguish it from wcpm that, in some literatures, is labeled oral reading fluency.

Reading Passages and Questions

A set of six reading passages and associated comprehension questions were developed to obtain measures

of rate, accuracy, prosody, wcpm, and passage comprehension. Comprehension questions were included as a second measure of comprehension and as a way to emphasize to students that they were not simply to read quickly; they were expected to attend to meaning. Following the administration guidelines used in most accountability systems, we used only grade-level passages for this study. All passages were drawn from trade books, magazines, and textbooks typical of those that students encounter in classrooms. Two passages, one narrative and one expository, at each of grades 2, 4, and 6 as estimated by Fry and Dale-Chall readability formulas, were administered to students at their assigned grade levels. Each passage was estimated to take a minimum of 3 minutes for average students to complete, and each contained sufficient content to construct five open-ended comprehension questions. Approximately half of the questions were designed to test high-level textually implicit understanding and half were designed to test textually explicit understanding of important information (Pearson & Johnson, 1978). The length of the passages ranged from an average of 361 words at grade 2 to 574 words at grade 6. In addition, none of the texts was picture dependent. All were judged to cover topics somewhat familiar to students and also contain information that was new or unfamiliar to guard against questions that could be answered correctly from prior knowledge.

All the passages and comprehension questions were reviewed by an independent, professional reading test developer who judged the passages to be grade and topic appropriate, and the questions to be clear and appropriately classified as textually explicit or implicit. In addition, all texts and questions were field-tested and revisions were made based on the pilot data. For purposes of administration, each passage was formatted into a test booklet with the title and a simple picture on the cover. The text was printed on both inside pages so that no page turning was required once testing began.

Procedure

The ITBS comprehension test was administered to students at the beginning of the school year. Trained administrators who were not aware of the students' scores on the ITBS test individually administered the oral reading assessment to students over two days from October to December. In the first session, students were introduced to the task and then read either the narrative or expository grade-level passage. In the second session, they read the passage from the other genre. The order of genre administration was counterbalanced.

Examiners advised students to read orally as they would read in class. If they came to a word they could not read, they were to try their best or skip it. No assistance was provided for unknown words. If students paused for 3 seconds, they were prompted to go on and were not supplied with any additional information or prompt so that passage comprehension was not compromised. Students were also told that after reading they would be asked several questions about the passage. As students read, the examiner noted the errors on a master copy of the passage. Immediately following the oral reading of each text, the passage was removed and students were asked to respond orally to five questions. All testing sessions were digitally recorded for later rescoring and interrater reliability checks. Two raters scored all the data for accuracy and time; two different raters scored for prosody and comprehension. All interrater reliability scores were calculated using the absolute agreement option of intraclass correlations (ICC; SPSS, 2004) that measures raters' absolute score agreement as well as ranking.

Scoring

Accuracy

Accuracy was scored by listening to digital recordings, reviewing the errors scored by examiners at the time of the interviews, and revising the scoring as necessary. Mispronunciations, substitutions, omissions, and words on which the student paused more than 3 seconds were scored as errors. Repeated words, self-corrections, words decoded slowly but ultimately read correctly, and mispronunciations due to dialect or regional differences were not counted as errors. Accuracy was calculated as the percentage errors of the total words read per minute; it was calculated at 1 minute and 3 minutes of oral reading. Interrater reliability was greater than 99%.

Rate

Rate was defined as the number of words students read per minute, unadjusted for errors. This measure provided an unambiguous measure of rate. As noted previously in the Procedure section, students who paused were allowed 3 seconds before they were prompted to go on. Rate scores were calculated as words per minute (wpm) for each student after 1 minute and 3 minutes of reading. Interrater reliability was greater than 99%.

Wcpm

The standard procedure for calculating wcpm was used—the number of words read in 1 minute minus the number of errors. Using the same procedure as described for scoring accuracy and rate, we scored mispronunciations, substitutions, omissions, and words on which the student paused more than 3 seconds as errors. Self-corrected errors, repeated words, and mispronunciations due to dialect or regional differences were not counted as errors. Students were prompted to

continue reading if they hesitated for 3 seconds. For this study, scores were calculated for 1 minute, the usual length of time for measuring wcpm. We also calculated wcpm for 3 minutes by averaging across 3 minutes of oral reading. Interrater reliability for both the 1-minute and 3-minute scoring was greater than 99%.

Prosody

Prosody was scored using the 4-point oral reading fluency scale adapted from NAEP, which focused on phrasing, adherence to the author's syntax, and expressiveness (Daane et al., 2005; Pinnell et al., 1995). A rubric was developed for each grade level (2, 4, 6), and examiners were trained using a strategic sample of audiotapes representing a range of student abilities within each grade. Following NAEP, scores of 1–2 on this scale were considered nonfluent, and scores of 3–4 were considered fluent at each grade level. Because 1 minute offered a very limited sample on which to score prosody, examiners listened to the audio recording of each student and assigned a score to each line of text read aloud over the first 3 minutes of reading. Then, they assigned an overall score to the section. Interrater reliability was 82%.

Passage Comprehension

A 3-point rubric was designed, pilot tested, and applied to score each comprehension question associated with the reading passages. In general, 0 indicated an incorrect response or no answer, 1 indicated a partially correct response, and 2 indicated a complete, correct response. If incorrect information was included along with the correct response, 1 point was deducted from the item score. Interrater reliability for comprehension questions was 94%.

Data Analysis

Two general decisions were made prior to data analysis. First, students' scores for each variable were calculated as the average performance of the grade-level expository and narrative passages (cf. Fuchs et al., 1988; Hosp & Fuchs, 2005). Such an average algorithm provided more reliable estimates of students' performance than those based on one passage. Accordingly, the data file included students who read both passages at their grade level. A series of *t*-tests on all passage-related dependent variables was run to compare the scores of students who read both passages with those who, due to absence, read one at each grade level.[1] No significant differences were found. Second, we were concerned about the possible statistical effects of restricted range of the distribution of accuracy scores. To determine if this was a problem, we replicated our data analyses using accuracy scores from a reverse log transformation and found that the results were generally identical to those with raw scores.[2] Given the similarity, we used percentage of errors instead of transformed scores for all analyses involving accuracy due to their inherent clarity for interpretation of results.

Results

Data analysis proceeded in three stages. First, we examined descriptive statistics and a series of correlations to investigate the relation among indicators of oral reading fluency (wcpm, accuracy, rate, prosody, comprehension) as well as the effect of reading for 3 minutes as compared with 1 minute. Next, we tested a set of models that could be used to assess oral reading fluency, exploring the relative contribution of wcpm, rate, accuracy, and prosody to reading comprehension across grade levels. Finally, we examined the degree to which a metric of wcpm misidentified students who were at risk for reading failure, and then qualitatively analyzed sample profiles of students who fell into various categories of risk, with an eye toward examining the appropriateness of using wcpm data to guide instruction.

Descriptive statistics for key variables by grade are presented in Table 1. Overall, the MANOVA analysis revealed a significant grade-level effect ($p < .001$) for all variables that were not standardized within grade level (i.e., wcpm—1 minute, wcpm—3 minutes, rate—1 minute, rate—3 minutes). The data also indicated that the sample of students in this study were fairly average as indicated by mean scores on several variables. Across the grades, ITBS scores placed students at approximately the 50th percentile, as did scores for wcpm when compared with Hasbrouck and Tindal (2006) benchmarks. In addition, mean accuracy and passage comprehension scores were consistent with scores commonly used by informal reading inventories to signify instructional level (Lipson & Wixson, 2009). Correlations between the ITBS comprehension scores and passage comprehension were .50, .59, and .63 for grades 2, 4, and 6 respectively, indicating a moderate relation between scores on the ITBS shorter passage, multiple-choice items and the scores from longer passage, open-ended items.

Similar to findings from the NAEP Oral Reading Studies (Daane et al., 2005; Pinnell et al., 1995), students demonstrated strong accuracy when reading grade-level texts (≥ 94%); although, unlike the NAEP studies, these students were seeing the passages for the first time and all passages were estimated to be proportionately longer than NAEP passages were for fourth grade. In general, mean scores for accuracy, rate, and wcpm decreased ($p < .001$) as students read for 3 minutes. Thus, students read less accurately and less quickly when reading for

slightly longer. The one exception was at sixth grade where there were no significant differences for rate or wcpm after 1 minute as compared with 3 minutes of reading, although there was a significant decrease for accuracy. The finding that 1 minute of reading overestimates reading rate for second- and fourth-grade students is consistent with the findings of the NAEP 2002 Oral Reading Study (Daane et al., 2005).

The next set of analyses was conducted to explore the relation between wcpm and rate, accuracy, prosody, and comprehension—indicators of oral reading fluency. Results of Pearson correlation analyses (see Table 2) indicated that wcpm is largely a measure of rate (r = .99), a finding consistent with Pressley and colleagues (2005). Although accuracy and rate are figured into the wcpm metric, accuracy contributed relatively little of the wcpm score variance. However, there was a strong and consistent correlation across grades between wcpm and prosody that is likely influenced by rate-related aspects of prosody (i.e., phrasing and flow) that are included in the prosody rubric.

Results also indicated that wcpm for 1 minute of reading was not as strongly correlated with reading comprehension on the ITBS as suggested by earlier research. The correlation between wcpm and ITBS comprehension for grades 2, 4, and 6 was .55, .48, and .48, respectively, accounting for 23%–30% of the variance. Correlations between wcpm and comprehension scores on the actual passages read as part of the assessment were significant at grades 2 and 4, although they were quite low, most especially at grade 6. As we noted in the Methods section, these assessment passages were

Table 1. Descriptive Statistics: Means and Standard Deviations by Grade

Variables		*N* = 93 Grade 2	*N* = 91 Grade 4	*N* = 95 Grade 6
Wcpm—1 minute	*M*	87.61	120.75	126.41
	SD	28.80	28.49	24.55
Wcpm—3 minutes	*M*	81.31	116.16	125.72
	SD	28.59	31.12	28.75
Rate (wpm)—1 minute	*M*	90.78	123.30	129.86
	SD	28.45	28.45	24.17
Rate (wpm)—3 minutes	*M*	85.10	119.60	129.52
	SD	28.37	31.31	28.40
Accuracy (% errors)—1 minute	*M*	.045	.025	.031
	SD	.039	.020	.020
Accuracy (% errors)—3 minutes	*M*	.059	.037	.036
	SD	.042	.024	.020
Prosody (1–4)	*M*	2.30	2.95	3.15
	SD	.76	.75	.60
Passage comprehension (% correct)	*M*	71.72	74.58	68.27
	SD	21.79	19.97	28.33
ITBS Comprehension (NCE)	*M*	49.72	57.04	51.61
	SD	15.43	17.88	16.20

Note. ITBS = Iowa Test of Basic Skills Battery; wcpm = words correct per minute; wpm = words per minute.

Table 2. Correlation of WCPM With Rate, Accuracy, Prosody, and Comprehension by Grade

Grade	Rate	Accuracy (errors)	Prosody	ITBS comprehension	Passage comprehension
2	.99***	–.46***	.84***	.55***	.24*
4	.99***	–.39***	.84***	.48***	.29**
6	.99***	–.43***	.84***	.48***	.12

Note. ITBS = Iowa Test of Basic Skills Battery.
*$p < .05$. **$p < .01$. ***$p < .001$.

longer and more conceptually complex than most found on norm-referenced tests, and 50% of the questions targeted higher-level comprehension.

The last set of correlations examined whether the relations of rate, accuracy, and wcpm with ITBS comprehension changed depending on the length of time students read aloud. We included calculations for wcpm at 3 minutes of reading to determine if simply extending the reading time improved the correlations with comprehension. Table 3 shows that at all grades, correlations were consistent across time for rate and wcpm with respect to ITBS comprehension. However, correlations between accuracy and ITBS comprehension increased at 3 minutes of reading for grades 2 and 4, suggesting that, in general, having students read for slightly longer may increase the predictive power of accuracy scores.

Taken together, the descriptive and correlational analyses informed the next set of analyses in two ways. First, differences among the strength of the correlations between the various indicators of oral reading fluency and comprehension suggested using separate variable scores rather than a composite metric of wcpm. In this way, results might provide a finer grained understanding of oral reading fluency and more specific directions for instruction. Second, based on multiple considerations, we decided to use 3-minute measures of rate, accuracy, and prosody for the multivariate regression analyses that follows. Specifically, we were persuaded by the significant differences in mean scores of rate and accuracy for 1 minute and 3 minutes of reading, the increased correlation between accuracy and ITBS at 3 minutes, the need to measure prosody after 3 minutes of reading, and NAEP findings that 1 minute underestimated errors and overestimated rate.

The second stage of analysis tested a series of multiple regression models using structural equation modeling (SEM) to analyze how rate, accuracy, and prosody predict ITBS comprehension scores. Maximum likelihood estimates of the standardized regression coefficients were obtained with the AMOS program (Arbuckle, 1997). Using the SEM procedure to conduct the multiple regression analyses enabled us to evaluate the unique variance accounted for by each of the variables after the shared variance was removed. In addition, this statistical procedure allowed examination of all the models simultaneously and the analysis of multiple linear regressions for the three grade levels. The program executes statistical comparisons of coefficients between and across grades, providing more efficient parameter estimates than estimation done separately with any single group model (Arbuckle & Wothke, 1995).

Four successive models were examined: (1) the baseline model of ITBS comprehension predicted by wcpm, (2) ITBS comprehension predicted by wcpm and prosody, (3) ITBS comprehension predicted by simultaneously considering prosody, rate, and accuracy, and (4) ITBS comprehension predicted by simultaneously considering prosody, rate, accuracy, and passage comprehension. The results include both the standardized regression coefficients (i.e., the correlation between a predictor and the dependent variable) and percentage of variance explained by all the predictors (i.e., the proportionate reduction in error in estimating the dependent variable, R^2; Pedhazur & Kerlinger, 1982). The AMOS program provided statistical tests of regression coefficients between and across grades. We also conducted F tests to statistically compare the four different models using the R^2 estimates.[3]

Figure 1 displays the relation between the single measure of wcpm and ITBS comprehension across grades. As indicated previously, wcpm scores based on 1 minute of reading accounted for only 30%, 23%, and 23% of the variance in comprehension scores, decreasing after grade 2. We also tested this model using scores of wcpm after 3 minutes of reading to determine if simply having students read for longer would increase the correlation between wcpm and comprehension. Results for wcpm measured at 3 minutes were not significantly different than those based on 1-minute measures for all grade levels (grade 2, $F(1, 91) = 3.73$, $p > .05$; grade 4, $F(1, 89) = 1.17$, $p > .05$; grade 6, $F(1, 93) = 2.48$, $p > .05$).

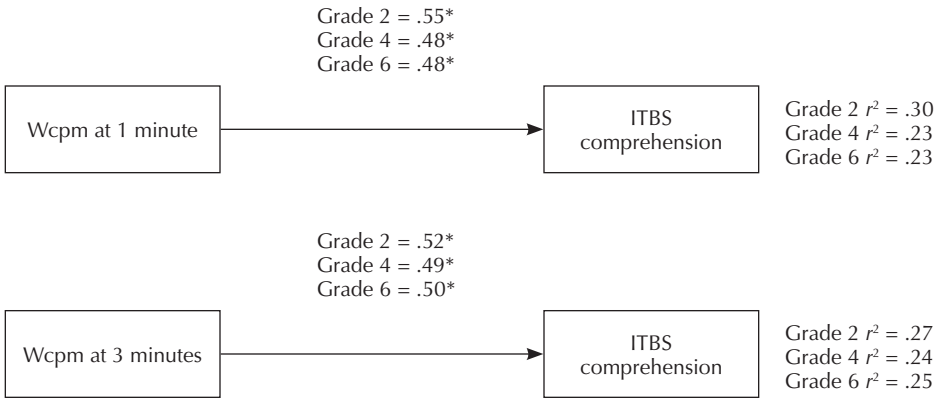
Next, we examined a model in which prosody was added to wcpm to predict comprehension (see Figure 2). Because the wcpm metric includes rate and accuracy,

Table 3. Correlations for 1-Minute and 3-Minute Measures of Rate, Accuracy, and Wcpm With ITBS Comprehension by Grade

	Rate 1 minute	Rate 3 minutes	Accuracy 1 minute	Accuracy 3 minutes	Wcpm 1 minute	Wcpm 3 minutes
Grade 2	.54**	.51**	–.37**	–.43**	.55**	.52**
Grade 4	.47**	.46**	–.35**	–.42**	.48**	.49**
Grade 6	.47**	.50**	–.24*	–.24*	.48**	.50**

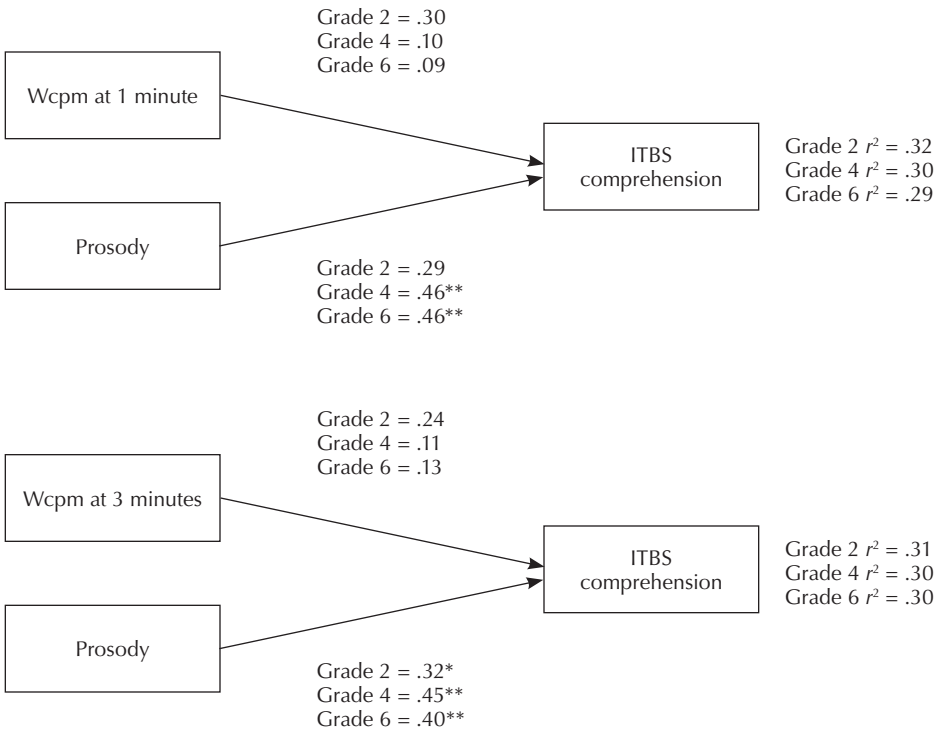
Note. ITBS = Iowa Test of Basic Skills Battery; wcpm = words correct per minute.
*$p < .05$. **$p < .001$.

Figure 1. Model for Predicting ITBS Comprehension From Wcpm at 1 Minute and 3 Minutes of Reading



Note. ITBS = Iowa Test of Basic Skills Battery; wcpm = words correct per minute.
*$p < .001$.

Figure 2. Model for Predicting ITBS Comprehension From Wcpm and Prosody at 1 Minute and 3 Minutes of Reading



Note. ITBS = Iowa Test of Basic Skills Battery; wcpm = words correct per minute.
*$p < .05$. **$p < .01$.

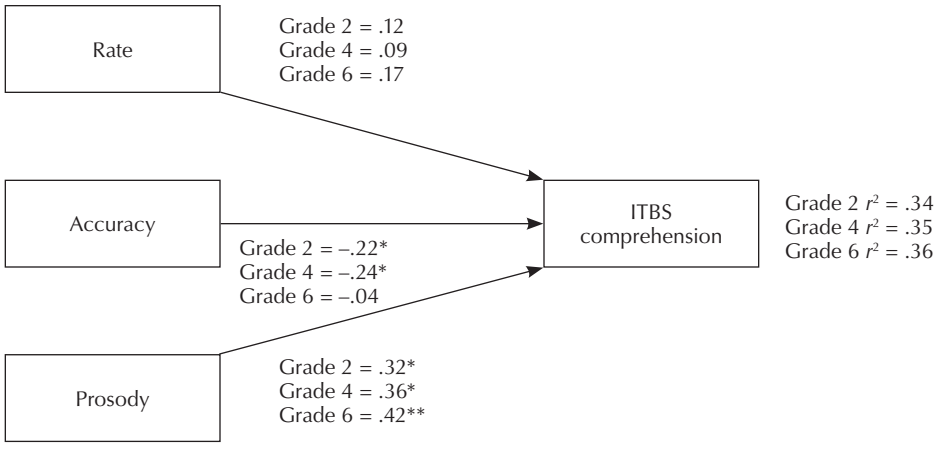
we wanted to test if adding a measure of prosody to this simple measure would enhance the correlation with comprehension. Here, again, we tested the model using wcpm scores for 1 minute and 3 minutes. We tested the 1-minute metric because that is typically used with wcpm; we tested the 3-minute measure of wcpm because the prosody measure was taken at 3 minutes and, as described previously, we believed the added

time might provide a more stable and authentic measure of oral reading fluency. These models using wcpm and prosody accounted for 29%–32% of the variance in comprehension scores depending on the grade level and whether measures were based on 1 minute or 3 minutes of reading. There were no statistically significant differences between the 1-minute and 3-minute models using wcpm and prosody to predict comprehension (grade 2, $F(1, 90) = 1.30$, $p > .05$; grade 4, $F(1, 88) = 0$, $p > .05$; grade 6, $F(1, 92) = 1.31$, $p > .05$). There were, however, statistically significant increases in amount of variance accounted for in all grades except grade 2 for 1 minute when prosody was added to the model as compared with the simple model using only wcpm (1 minute—grade 2, $F(1, 90) = 2.65$ $p > .05$; grade 4, $F(1, 88) = 8.80$, $p < .01$; grade 6, $F(1, 92) = 7.77$, $p < .01$; 3 minutes—grade 2, $F(1, 90) = 5.22$, $p < .05$); grade 4, $F(1, 88) = 7.54$, $p < .01$; grade 6, $F(1, 92) = 6.57$, $p < .01$). In addition, the relative contributions of wcpm and prosody changed across grades. Prosody and wcpm contributed similarly to comprehension at grade 2 but at both grade 4 and 6, prosody contributed significantly more to comprehension than wcpm (1 minute—grade 2, $z = 1.67$ $p > .05$; grade 4, $z = 2.74$, $p < .01$; grade 6, $z = 2.87$, $p < .01$; 3 minutes—grade 2, $z = 1.93$, $p > .05$; grade 4, $z = 2.70$, $p < .01$; grade 6, $z = 2.58$, $p < .01$).

The next model (see Figure 3) includes three separate variables associated with oral reading fluency—rate, accuracy, and prosody. This model was tested to isolate the contribution of three core indicators of fluency as compared with the previous model in which rate and accuracy were confounded in the wcpm metric. It allowed us to examine the relative contribution of each indicator, measured at 3 minutes, and to compare that contribution across developmental levels. Overall, this model accounted for more variance in ITBS comprehension scores (34%, 35%, 36%) than wcpm alone or wcpm and prosody together. There were statistically significant increases for all grades over the traditional 1-minute wcpm model in Figure 1 (grade 2, $F(2, 89) = 5.39$, $p < .01$; grade 4, $F(2, 87) = 16.06$, $p < .0001$; grade 6, $F(2, 91) = 18.48$, $p < .0001$) and over the 3-minute two-indicator model in Figure 2 for all grades (grade 2, $F(1, 89) = 4.05$, $p < .05$; grade 4, $F(1, 87) = 6.71$, $p < .01$; grade 6, $F(1, 91) = 8.53$, $p < .01$). From a developmental perspective, this three-indicator model added more explanatory power to comprehension over 1-minute wcpm at the intermediate grades than for grade 2 (the increases were as follows: grade 2 = 4%, grade 4 = 12%, grade 6 = 13%). There were also trends toward an increasing contribution of prosody across all the grades, and a decreasing contribution of accuracy between grades 4 and 6. These trends seem to be consistent with research on the decreasing role of decoding and increasing relation between prosody and comprehension as students move to higher levels of reading and comprehension.

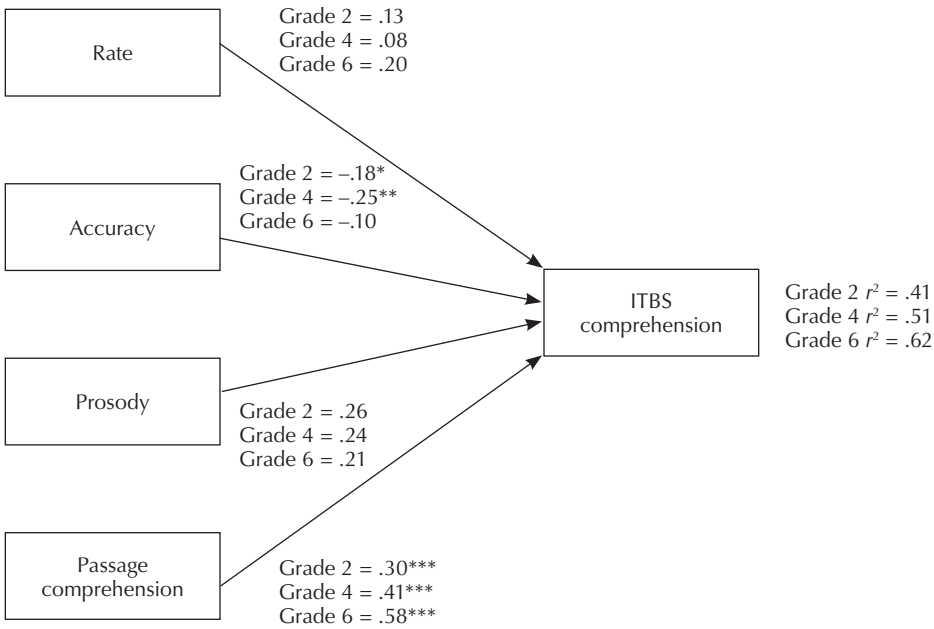
The fourth model (see Figure 4) examined the predictive power of adding passage comprehension to measures of rate, accuracy, and prosody. In keeping with a simultaneous focus on comprehension and oral reading fluency, we designed measures that oriented students to comprehension before reading and had them answer questions after reading. Therefore, we were able to test a model that included passage comprehension as

Figure 3. Model for Predicting ITBS Comprehension From Rate, Accuracy, and Prosody by Grade



Note. ITBS = Iowa Test of Basic Skills Battery; wcpm = words correct per minute.
*$p < .05$. **$p < .01$.

Figure 4. Model for Predicting ITBS Comprehension From Rate, Accuracy, Prosody, and Passage Comprehension by Grade



Note. ITBS = Iowa Test of Basic Skills Battery.
*p < .05. **p < .01. ***p < .001.

a predictor of ITBS total comprehension. Rather than considering comprehension simply as a criterion outcome, we were interested in the effect of holding students accountable for comprehension of the actual text read during assessment and how this variable might contribute to the predictive power of the model. Thus, we positioned passage comprehension as an integral component of the fluency assessment model.

When passage comprehension was added to the model that included separate measures of accuracy, rate, and prosody, the amount of variance accounted for in ITBS comprehension scores increased significantly over the model without text comprehension (Figure 3) for every grade level (the increases were as follows: grade 2 = 7%, $F(1, 88) = 10.44$, $p < .01$; grade 4 = 16%, $F(1, 86) = 28.08$, $p < .0001$; grade 6 = 26%, $F(1, 90) = 61.58$, $p < .0001$). A combination of rate, accuracy, prosody, and passage comprehension accounted for 41% of the variance at grade 2, 51% of the variance at grade 4, and 62% of the variance at grade 6. As with the previous models, developmental trends were seen, marked by an increasing contribution of comprehension as the grade level increased and decreasing role of accuracy after grade 4. With the large contribution of passage comprehension and its shared variance with other variables, few of them significantly contributed additional unique variance to ITBS comprehension.

The third and final stage of analysis investigated the consequential validity of using a measure of wcpm to screen students at risk for reading difficulty (i.e., performing below grade level). Specifically we were interested in how accurately two commonly used wcpm benchmark scores (DIBELS; Hasbrouck & Tindal, 2006) identified students who were reading below grade level and what information teachers might find helpful in planning appropriate instruction for individual students.

We used a two-stage process to determine if wcpm benchmarks might yield either false positives (i.e., identify students as at risk who are not at risk) or false negatives (i.e., fail to identify students as at risk who are at risk). First, we examined the ITBS comprehension scores of students whose wcpm scores placed them in one of three DIBELS categories (low risk, some risk, at risk) and in one of four Hasbrouck and Tindal categories (0–24th percentile, 25th–49th percentile, 50th–74th percentile, 75th–99th percentile) for the appropriate time of year. Results, calculated as the percentage of students within each category, are displayed in Tables 4 and 5.

Following Schilling and colleagues (2007), we examined false positives and negatives by focusing on students in the low-risk and at-risk DIBELS categories and on students in the 0–25th percentile and 75th–99th

Table 4. Percentage of False Negatives and False Positives Using DIBELS Benchmark Scores and Categories as Indicated by ITBS Comprehension Scores

DIBELS category	ITBS > 60th percentile	ITBS < 25th percentile	ITBS < 40th percentile
Low risk			Grade 2 = 24% Grade 4 = 12% Grade 6 = 24%
Some risk		Grade 2 = 28% Grade 4 = 19% Grade 6 = 25%	
At risk	Grade 2 = 5% Grade 4 = 17% Grade 6 = 21%		

Note. ITBS = Iowa Test of Basic Skills Battery.

Table 5. Percentage of False Negatives and False Positives Using Hasbrouck and Tindal Benchmark Scores and Categories as Indicated by ITBS Comprehension Scores

Hasbrouck and Tindal[a] category	ITBS > 50 percentile	ITBS > 75 percentile	ITBS < 25 percentile	ITBS < 50 percentile
75th–99th percentile				Grade 2 = 27% Grade 4 = 14% Grade 6 = 20%
50th–74th percentile			Grade 2 = 7% Grade 4 = 0% Grade 6 = 7%	
25th–49th percentile		Grade 2 = 7% Grade 4 = 35% Grade 6 = 22%		
0–24th percentile	Grade 2 = 0% Grade 4 = 27% Grade 6 = 24%			

Note. ITBS = Iowa Test of Basic Skills Battery.
[a]Hasbrouck, J., & Tindal, G.A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher, 59*(7), 636–644.

percentile Hasbrouck and Tindal categories. False positives were identified for 5%–21% of the students identified as at risk according to DIBELS scores and 0%–27% of the students identified in the 0–24th percentile by Hasbrouck and Tindal. These "at-risk" students scored above grade-level norms on the ITBS comprehension test despite having very low wcpm scores. At grade 2, few, if any, students with low wcpm scores scored above grade level on the ITBS comprehension test; however, a substantial portion of students in grade 4 and 6 read slowly or inaccurately enough according to wcpm measures to be considered at risk, yet they had strong comprehension. Developmental reading models would, again, suggest that rate and accuracy lose predictive power as reading comprehension becomes more sophisticated and texts become more complex. As would be expected, rate and accuracy play a bigger role in comprehension at the early stages of reading.

Benchmark scores for both DIBELS and Hasbrouck and Tindal (2006) were also used to identify false negatives—students with high wcpm scores identified as "low risk" according to DIBELS or in the 75th–99th percentile according to Hasbrouck and Tindal who also scored below grade-level expectations on the ITBS comprehension. False negatives are often of more concern to educators than false positives because failure to identify students in need results in lost opportunities for intervention and possible compounding of reading difficulties over time (Jenkins, Hudson, & Johnson, 2007). Specifically, 12%–24% of students identified by DIBELS and 14%–27% identified by Hasbrouck and Tindal could be classified as false negatives. Taking a more liberal approach to identifying false negatives,

that is, considering students who were categorized in the 50th–74th percentile or at some risk according to wcpm scores yet scored below the 25th percentile on ITBS comprehension, the number of false negatives increases. These students would not be identified for intervention based on their wcpm scores although they were actually at risk based on ITBS scores; overall, they read so quickly and accurately that they were not anticipated to have reading difficulty yet their comprehension scores were below grade level.

Another procedure for estimating how well a measure discriminates students who are truly at risk is to examine a sensitivity index that is calculated as the ratio of true positives to the sum of true positives and false negatives. A test that perfectly identifies students at risk would have 100% sensitivity; an acceptable level of sensitivity for screening tests is recommended to be 90%–95% (Johnson et al., 2009). Calculations of the sensitivity of wcpm measures using DIBELS categories produced the following: grade 2 = 77%, grade 4 = 78%, grade 6 = 66%. Calculations using Hasbrouck and Tindal categories produced the following: grade 2 = 79%, grade 4 = 93%, grade 6 = 79%. Across both classification systems and all grade levels, five of the six calculations indicated unacceptable levels of sensitivity. In general, false negatives were less pronounced in our data at grade 4 than at either grade 2 or 6, a finding that may reflect the slightly higher performance on ITBS comprehension of the grade 4 sample than those at the other grades (see Table 1).

As a second approach to examining consequential validity and implications for instruction, we examined the reading profiles of individual students by analyzing the interplay of accuracy, rate, prosody, and comprehension. Based on our finding that rate, accuracy, and prosody contributed differentially to predictions of comprehension, we wondered if a measure of wcpm might mask students' individual strengths and weaknesses that could be important targets for instruction. Table 6 depicts sample profiles of students who would be classified according to the DIBELS and Hasbrouck and Tindal (2006) categories based on their wcpm scores, along with their scores on each of the indicator variables. Using percentages associated with informal reading inventories and classroom assessments (Lipson & Wixson, 2009) the following scores guided our interpretation of these profiles: accuracy above 95% = independent level, 90%–95% = instructional level, below 90% = frustration level; rate scores indicating at-risk performance for grade 2 < 57, grade 4 < 91, grade 6 < 109; prosody scores of 1–2 = not fluent, 3–4 = fluent; ITBS comprehension less than 40th percentile = below grade level, above 60th percentile = above grade level; passage comprehension below 50% = frustration level, above 70% = instructional/independent level.

For example, the two second-grade students whose wcpm scores placed them in the 0–24th percentile category according to Hasbrouck and Tindal (2006) benchmarks and at risk according to DIBELS benchmarks displayed different profiles of strengths and weaknesses even though their wcpm scores were quite similar. Both students scored in the lowest range for rate and prosody yet one student was at frustration level for decoding (82% accuracy) and the other was at instructional level (94%). Their ITBS comprehension scores were similar, but their performance on inferential passage comprehension questions was markedly different. Analysis of student profiles such as these provides diagnostic insight into students' skills that are not evident from their wcpm scores.

We analyzed the profiles of all participants in low performing wcpm groups (i.e., at risk for DIBELS; 0–24th percentile for Hasbrouck and Tindal, 2006) and in the high performing wcpm groups (i.e., low risk for DIBELS; 75th–99th percentile for Hasbrouck and Tindal) to determine the percentage of students in each of these categories who exhibited specific areas of reading difficulty that might inform instruction (Table 7). Students were considered to have demonstrated difficulty if their profiles indicated below acceptable scores in accuracy, rate, prosody, or comprehension as described previously. Consistent with the findings for false negatives, the majority of students misidentified as low risk exhibited difficulty with comprehension although prosody and accuracy were also problem areas for a portion of grade 2 students. It is not surprising that few students in this category exhibited difficulty with rate or accuracy because their wcpm scores were high, placing them at low risk. Although we expected to find areas of need for the students classified as at risk, the profiles of these students suggested varying patterns that would require different instructional emphases. In general, decoding was an area of strength for the students with low wcpm scores, but they demonstrated considerable difficulty across the other three indicators of oral reading fluency—rate, prosody, and comprehension. Overall, the profile descriptions and findings of false negatives and false positives for wcpm reinforce the importance of examining the full range of variables that underlie skilled oral reading fluency.

Discussion

This study raises issues about the widespread use of wcpm metrics to assess oral reading fluency and to identify students at risk of reading difficulty. Overall, we found that a wcpm score taken after 1 minute of reading is largely a measure of rate. When separate indicators of oral reading fluency (rate, accuracy, prosody,

Table 6. Sample Qualitative Profiles of Students by Grade

Grade	Hasbrouck and Tindal[a] category	Sample Student	DIBELS Category	Wcpm	Accuracy (% correct)	Rate (Wpm)	Prosody (1–4)	Passage comprehension (%)	ITBS (NCE)
2	> 75th percentile	1	Low risk	124	98	127	3	42	38
		2	Low risk	168	99	169	4	67	80
	50th–74th percentile	3	Low risk	90	87	102	1	42	41
		4	Low risk	100	97	103	2	33	39
	25th–49th percentile	5	Some risk	73	97	75	3	83	80
		6	At risk	55	89	62	2	83	47
	0–24th percentile	7	At risk	44	82	54	2	33	34
		8	At risk	47	94	53	1	67	38
4	> 75th percentile	1	Low risk	151	89	169	3	58	70
		2	Low risk	147	99	142	4	25	48
	50th–74th percentile	3	Low risk	118	94	126	2	33	42
		4	Low risk	123	93	137	3	100	64
	25th–49th percentile	5	Some risk	93	93	97	2	92	45
		6	Some risk	97	96	99	3	33	36
	0–24th percentile	7	At risk	88	94	94	2	92	66
		8	At risk	45	96	47	1	33	31
6	> 75th percentile	1	Low risk	175	100	175	4	17	39
		2	Low risk	185	96	190	4	100	88
	50th–74th percentile	3	Low risk	154	100	154	4	58	70
		4	Low risk	164	99	164	3	50	46
	25th–49th percentile	5	Low risk	122	98	130	2	0	20
		6	Some risk	119	92	126	3	17	57
	0–24th percentile	7	At risk	73	99	74	3	90	59
		8	At risk	102	92	118	2	50	29

Note. wcpm = words correct per minute; wpm = words per minute.
[a]Hasbrouck, J., & Tindal, G.A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher, 59*(7), 636–644.

Table 7. Percentage of Students Exhibiting Below-Grade-Level Performance in Components of Oral Reading Fluency by Grade and Risk Group

	Students classified as low risk				Students classified as at risk			
Grade	Accuracy	Rate	Prosody	Comp	Accuracy	Rate	Prosody	Comp
2	15%	0%	37%	37%	14%	62%	100%	67%
4	0%	0%	7%	31%	5%	72%	92%	80%
6	0%	0%	2%	69%	5%	75%	49%	56%

passage comprehension) were used in assessment, the result provided a finer grained understanding of oral reading fluency and fluency assessment, and a stronger predictor of comprehension. In this discussion, first we turn to issues of construct validity—the definition and assessment of oral reading fluency. The issues of criterion and consequential validity are considered later.

Construct Validity

This study acknowledges the history and development of wcpm as a general indicator of reading ability and also calls into question its prevalence as a measure of oral reading fluency, especially in current high-stakes contexts where instruction is likely to mirror assessment. Unfortunately, Deno and colleagues' (1982, 1985) early cautions about the risks of teaching to a test of wcpm and the potential for comprehension to become detached from reading aloud seem to have been borne out with increasing frequency (Newman, 2009; Pearson, 2006; Pressley et al., 2005). Definitions of oral reading fluency and the interactive nature of the reading process would seem to recommend assessments that consider multiple facets of oral reading—the combined role of rate, accuracy, and prosody in contributing to comprehension. We tested several models of oral fluency assessment that included these multiple indicators and that align with Samuels' (2006) call for oral reading fluency assessments that mimic fluent reading by demanding simultaneous decoding and comprehension. Specifically, we oriented students to read for understanding before they began the assessment; used longer, more complex texts to engage comprehension; and then followed up with textually explicit and implicit questions.

Comparing various models for assessing oral reading fluency, we found that a model composed of separate measures of rate, accuracy, and prosody accounted for a statistically significant increase in variance in comprehension scores across all grades compared with a single measure of wcpm or even a model that added prosody to wcpm. This finding is consistent with Rasinski's (1990) three-factor model of oral reading fluency and NAEP's attention to all three variables in their study of oral reading (Daane et al., 2005; Pinnell et al., 1995). Although adding a prosody measure to a single wcpm measure did improve predictions of comprehension for grades 4 and 6, the model that included individual measures of rate, accuracy, and prosody improved the prediction even further for all grades. This model confirmed differential contributions and developmental shifts in indicators of oral reading fluency that have also been posited by others (Fuchs et al., 2001; Paris et al., 2005; Schwanenflugel et al., 2006). In particular, the increasing contribution of prosody across the grades and the decreasing contribution of accuracy after grade 4 align with Chall's (1996) developmental model of reading acquisition. In that model, readers at Stage 2 (grades 2–3) develop increasing sophistication with decoding and automaticity, and at Stage 3 (grades 4–8) concentrate less on print and more on ideas and learning from more complex text. Our three-factor assessment model also confirmed the growing importance of prosody as an indicator of oral reading fluency and a strong correlate of comprehension at all grades but most especially at later stages of reading development.

The four-factor assessment model that added passage comprehension to accuracy, rate, and prosody accounted for significantly more variance in ITBS comprehension than the three-factor model. We concede that adding a passage comprehension indicator to predict more general comprehension is somewhat unconventional, yet we were persuaded to do so by two arguments: (1) the ecological validity of holding students accountable for comprehension as part of an oral reading fluency assessment and (2) the possibility that passage comprehension scores might enhance the criterion validity of oral reading fluency measures with more general comprehension measures. With the addition of comprehension scores from the assessment passages, the model resulted in large and statistically significant increases in the amount of variance accounted for at every grade over the assessment model without passage comprehension. It contributed greater explanatory power as grade level increased, a finding that aligns with studies suggesting a separate comprehension factor for oral reading fluency at the intermediate grades (Cramer & Rosenfield, 2008; Shinn et al., 1992). As students move through the grades and become more skilled readers, comprehension may have less to do with rate and accuracy and more to do with other indicators of oral reading fluency—namely higher-level comprehension (Figure 4) and prosody (Figure 3).

Although the three- and four-factor models in this study did not account for substantially more variance in comprehension than some prior studies that have simply used a single 1-minute wcpm measure (e.g., Good & Jefferson, 1998; Marston, 1989), they did add significantly beyond the 1-minute wcpm measure used in this study. More important, we believe these multicomponent models offer more valid measures of the fluency construct, important insights into the complex nature of fluency, informative developmental patterns, and diagnostic information to inform assessment and instruction. We address implications of this emphasis on the construct of fluency for assessment and instruction next.

First, our findings suggest that it is both feasible and desirable to expand measures of oral reading fluency to include core indicators of the construct (i.e., rate, accuracy, prosody, and perhaps passage comprehension).

Although assessing prosody is relatively more difficult than assessing rate and accuracy, the use of a holistic rubric resulted in reasonably strong reliability, and the addition of prosody proved to be informative at all grade levels.

Second, aligning the construct of oral reading fluency with a model of assessment makes the concept more transparent for both students and teachers, and it may guard against inappropriate instruction and test preparation. It may also support instruction that simultaneously targets all aspects of oral reading fluency while providing adaptations for individual differences (Rasinski, 2006; Rasinski & Padak, 2008), although the results of such interventions would need to be studied.

Third, the overall results and student performance profiles from an oral reading fluency assessment that includes separate measures of rate, accuracy, prosody, and comprehension would add specificity and diagnostic information that are necessary for effective instructional interventions. For example, teachers would be able to distinguish students who are strong in accuracy but weak in rate from those who are weak in accuracy but read quickly without monitoring for comprehension, or those who are automatic decoders yet read with little phrasing, expression, or understanding.

Fourth, both of the models tested in this study that included separate measures of oral reading fluency indicators (see Figures 3 and 4) support developmental trends for the oral reading fluency–comprehension connection and among the components of oral reading fluency. Specifically, although studies using measures of wcpm have suggested a decreasing relationship between fluency, as defined by wcpm, and comprehension as grade level increases, our data suggest this may be an artifact of the wcpm metric. The findings in this study indicate an increasing relationship between fluency and comprehension when fluency is defined as simultaneous attention to rate, accuracy, prosody, and comprehension. This suggests that, in general, oral reading fluency instruction in the early grades might well focus on rate, accuracy, prosody, and comprehension, and at higher grades or reading levels it might focus more on prosody and comprehension. Overall, developmentally different models of assessment and instruction should be explored (Jenkins et al., 2007; Paris et al., 2005). For example, assessments of oral reading fluency might place differential weighting on rate, accuracy, and prosody at different grade levels, or they might exclude or include specific indicators at particular grades or reading levels.

Finally, we are encouraged that the approach to assessing oral reading fluency used in this study is quite similar to many of the informal reading inventories and leveled reading passage assessments used in many schools. Consequently, it may be possible to enhance these assessments so they can provide additional information without introducing an entirely new form of assessment. In this study, students were given a comprehension focus for reading, read aloud from one complete expository and one narrative passage, answered textually explicit and implicit questions on the entire passage, and were scored for rate, accuracy, and prosody based on 3 minutes of oral reading averaged across two passages. Further research might explore having students read a portion of these longer texts aloud (approximately 3 minutes) and then complete the reading silently before answering comprehension questions. We found the average time to read each text aloud in its entirety was approximately 4 minutes for all grade levels, although the range of times was quite large. Since the scoring for all variables was taken at the 3-minute mark, it may not be necessary to extend oral reading beyond this point, yet the longer texts could remain the basis for assessing higher-level comprehension of the passage. This approach might also enable a comparison of oral and silent reading comprehension and rate that might be informative, particularly at the intermediate grades and above. Although the time to administer such assessments is more than the 3 minutes needed to administer three 1-minute timed readings, we believe that alignment with the fluency construct and the quality of information provided by such assessments make the process worthwhile.

As we investigated these alternative models for assessing oral reading fluency and issues of construct validity, we were mindful of keeping comprehension central to reading and of the various ways reading researchers and tests of oral reading fluency have positioned comprehension. By far, most have examined comprehension as an outcome measure, correlating wcpm or other measures of oral reading fluency with independent measures of comprehension, either concurrent or predictive. This was the approach we took here. However, some have suggested treating comprehension scores from the actual passages students read aloud as an outcome measure (Paris et al., 2005; Valencia et al., 2006). In this regard, our analysis confirms other studies that have demonstrated very low correlations between test passage comprehension and wcpm (Cramer & Rosenfield, 2008; Paris et al., 2005; Pressley et al., 2005).

Two possibilities may help to explain these low correlations between oral reading fluency indicators and test passage comprehension in this study and others, even when students are directed to read for meaning. First, the act of reading aloud may interfere with comprehension as readers focus on correct reading and performance over understanding. This is especially likely as grade and reading ability increase, and silent reading is the expected mode. Readers may be less likely to

engage in their usual comprehension strategies such as monitoring, fix-up, and problem solving under the pressure of reading aloud (Paris et al., 2005). Second, and most important in our view with respect to oral reading fluency, is that one can read quickly and accurately—and even with some prosody—yet not understand, most especially at the higher levels of comprehension. Conversely, when reading to understand complex text, good readers adjust their rates, slowing to monitor, think, reread, and such to understand. As our reader profiles demonstrated, oral reading fluency is situational and strategic when readers are attending to meaning and decoding (Samuels, 2006). Overall, having a comprehension focus for oral reading fluency assessments seems essential, in our view, to the construct validity of oral reading fluency and the goal of skilled reading. Under these conditions, the indicators of rate, accuracy, and prosody may not be strongly correlated with test passage comprehension, but they are good indicators of oral reading fluency. Test passage comprehension serves as an important and significant predictor of a students' overall comprehension. These issues of passage comprehension and oral reading fluency indicators should be the focus of future research.

Criterion and Consequential Validity

We turn next in our discussion to issues of criterion and consequential validity—using wcpm measures to identify students at risk for reading difficulty. Measures of 1-minute wcpm have become a popular strategy for screening students at risk of reading difficulty, largely based on their correlations, concurrent or predictive, with norm-referenced or state tests of comprehension. Overall, we found significant, although smaller, correlations at all grade levels between wcpm and ITBS comprehension scores than reported by earlier studies of wcpm or test developers. We could account for only 23%–30% of the variance in comprehension scores. This is somewhat higher than results reported by some (e.g., Pressley et al., 2005) but considerably lower than the average 64% of variance reported in much of the CBM literature (Good & Jefferson, 1998; Marston, 1989).

Our finding of lower correlations between wcpm and comprehension may have been influenced by the procedures, materials, or analyses used in this study. Specifically, we directed students to read for meaning. They were not aware they were being timed nor were there any time limits to complete reading the passages. It is possible that when good readers see comprehension as the goal, they are apt to slow down and engage in more self-monitoring than lower ability students (cf. Pressley et al., 2005) thereby decreasing the overall correlation of wcpm and comprehension. In contrast, when students are oriented to read for speed (as in most wcpm measures), the overall correlation between wcpm and comprehension may increase because good readers are able to speed up but poor readers are not. These differences in orientation to the fluency task may, in part, be responsible for the lower correlations between wcpm and comprehension in this study; conversely, they may be responsible for higher correlations in studies that emphasize speed rather than comprehension in oral reading fluency. Our choice of complex, longer texts and emphasis on high-level, open-ended questions may have also resulted in lower correlations between wcpm and comprehension as compared with the shorter texts and multiple-choice format of most standardized reading tests. Furthermore, the fact that we analyzed data within grade level rather than across grades and excluded special education students who were receiving supplemental services and beginning ELLs may have suppressed the correlations as compared with most wcpm studies that include a wider range of student abilities and grades (Mehrens & Clarizio, 1993). From the perspective of construct validity, however, we believe these decisions to orient students to meaning, focus on higher-level comprehension, and examine variability within and across grades were necessary to obtain a valid measure of oral reading fluency.

Most troubling with respect to consequential validity was the high rate of false negatives identified by wcpm measures when applying commonly used benchmarks (i.e., DIBELS; Hasbrouck & Tindal, 2006) and calculations of test sensitivity (Johnson et al., 2009). A substantial percentage of students across all grades who scored below grade level on the ITBS comprehension test were not identified as at risk of reading difficulty using wcpm scores derived from 1 minute of oral reading. These students would not have been identified or received intervention based on their scores. Our findings of under-identification parallel several other studies of screening accuracy using wcpm oral reading measures in which false negatives rates ranged from 15% to as high as 47%, depending on the benchmarks used (Jenkins et al., 2007; Pressley et al., 2005; Riedel, 2007; Schilling et al., 2007). The reader profiles of students in this study who were misidentified according to DIBELS and Hasbrouck and Tindal wcpm standards indicated that 30%–70% of the students demonstrated difficulty with comprehension. From the perspective of consequential validity, this finding is particularly disturbing both because comprehension is the sine qua non of reading and because a majority of low-performing ELLs struggle with comprehension (Lesaux, 2006). Potential problems associated with consequential validity are further highlighted by the various profiles of students who were correctly identified as at risk. If instructional interventions were focused on improving wcpm scores, our data suggest that more than 50% of the at-risk students

in this sample might not receive the comprehension or prosody instruction they need.

Findings related to false negatives and positives are inherently tied to benchmarks or cut scores used to classify students in various ways as at risk or not at risk. It is important to note, however, that much of the research on CBMs has involved developing local wcpm norms and monitoring students' growth over time, not applying predetermined benchmarks to identify students at risk (Deno & Marston, 2006; Stage & Jacobsen, 2001), and some of these same researchers have raised concerns about using common benchmarks. In this study, we applied both DIBELS and Hasbrouck and Tindal (2006) criteria because teachers, schools, districts, and states frequently use them, and because these scores have been integrated into many different instructional programs and generic guidelines for assessment, making the consequences significant for students.

Our findings regarding criterion and consequential validity of using wcpm measures have implications for assessment and instruction. We agree with others who caution that accurately screening students at risk requires careful consideration of multiple variables (e.g., reading levels of the students, text difficulty, criterion measure, etc.) and is best accomplished using multiple measures, including comprehension measures (Cramer & Rosenfield, 2008; Fuchs et al., 2001; Jenkins et al., 2007; Johnson et al., 2009; O'Connor & Jenkins, 1999; Schilling et al., 2007). However, in the rush to find quick and easy screening measures and meet the demands of federal and state policies, many states and school systems have not taken this path. Instead, they have chosen to rely solely on wcpm measures that may fail to identify a substantial number of students who have reading difficulty, most especially those with difficulty in comprehension. Therefore, we suggest that if screening measures are used, they must be designed to assess multiple aspects of reading and to accurately identify those who have needs in areas other than decoding and automaticity. If wcpm measures are used, they should be supplemented with comprehension assessments. In response to the concern about underidentification, some have suggested using a two-layer assessment system—using wcpm measures as a first-level screen and then moving to more in-depth diagnostic assessment for those identified by the first screen. Although in theory this seems like a logical sequence for assessment, the problem is that the first-level wcpm screen will miss too many students in need and fail to move them to the next level of assessment.

The various reader profiles of participants in this study and other studies (Fuchs, Fuchs, & Hamlett, 2007; Hosp & Fuchs, 2005; Valencia, in press) should also give us pause about using data from wcpm assessments to plan instruction or target intervention programs. Clearly, wcpm scores do not provide the depth or breadth of information needed to make good instructional decisions or programmatic recommendations. The wcpm metric itself obscures the relationship between accuracy and automaticity, and it fails to capture other critical aspects of skilled reading. Ultimately, decisions made on such limited data are destined to waste time and money, fail to achieve the larger goal of improving student learning, and fall short in meeting the needs of the students most at risk. Schools and teachers need to put wcpm data in perspective and to base programmatic and instructional directions on more complete assessments of students' needs.

Limitations

Several limitations should be considered in interpreting the results of this study and conducting future research. Methodologically, our choices related to reading passages (e.g., genre, length), orientation to read for meaning, assessment of higher-level comprehension, use of a prosody rubric, and longer oral reading times were selected to operationalize the construct of oral reading fluency and to engage students in decoding and comprehension simultaneously, yet these choices may limit the generalizability of our findings. Although studies of CBMs suggest that generic grade-level passages produce valid results (Fuchs & Deno, 1994), other researchers have raised issues about passage effects (e.g., Francis et al., 2008). Also, the addition of comprehension questions makes passage selection somewhat more complex. We are eager, therefore, for other researchers to investigate our reading fluency assessment models as well as other models using similar methods and different reading selections with an eye toward developmental differences.

We investigated the accuracy of using wcpm data to identify students at risk and examined reader profiles because of the connection to oral reading fluency and widespread use of wcpm measures. We did not set out to develop screening measures or to compare them, nor did we use standard passages or procedures typically found in specific wcpm tests. Thus, our results with respect to false negatives and false positives should be viewed with caution and replicated with other measures and samples. Nevertheless, our results suggest that a good deal more research will be needed if screening measures are to accurately identify students at risk according to clearly defined and valued outcome criteria. We also did not investigate the use of wcpm to monitor student progress over time, which has been the focus of a good deal of CBM-related work. Our findings may have implications for progress monitoring with respect to oral reading indicators and developmental shifts, but these issues will need to be the target of future studies. In our view, it would be particularly interesting to

study other progress monitoring models that include direct measures of comprehension or indicators of oral reading fluency, especially at the intermediate grades. Finally, although our student sample was drawn from seven schools across two districts, we were not able to document students' exposure to specific instructional approaches or interventions that may have influenced their performance in this study.

This study adds to a growing body of research on oral reading fluency, both its definition and assessment. As Samuels (2006) noted, "Definitions are not trivial, ivory tower concerns that are of no importance—how we define a construct such as fluency determines and influences to a large degree how we will measure it" (p. 39). Because assessment drives instruction—in best cases by modeling the construct to be learned and in worse cases by pressuring schools to teach whatever students need to score well—measures of oral reading fluency should meet the test of construct, criterion, and consequential validity. Our efforts here were designed to explore possible alternatives to wcpm measures that are transparent, lead to sound instructional decisions, and have better consequences for students. Our results point to new possibilities, and they suggest that caution is warranted as wcpm data are interpreted and used in school and policy contexts.

Notes

The research reported here was supported in part by the Spencer Foundation, Carnegie Corporation of New York, and the University of Washington College of Education. The opinions expressed are those of the authors and do not represent views of the funding agencies. We wish to thank Robert Abbott for statistical advice and Harry Katz for developing data retrieval programs for this study. In addition, we thank Joe Jenkins, Roxanne Hudson, and Deborah McCutchen for helpful feedback on earlier drafts of this manuscript.

[1] Power analysis was conducted for the proposed statistical analyses. The analysis revealed that for nearly all the analyses, inclusion of students who read only one passage did not change statistical power. Therefore, we decided to exclude those students in our data analyses instead of employing data imputation procedures to include those students.

[2] The log transformation was able to reshape the accuracy scores into a normal distribution but did not impact their relationship with other performance scores.

[3] $F = \frac{R^2 - R^2_{reduced}}{k - k_{reduced}} \div \frac{1 - R^2}{N - k - 1}$, where R^2 and $R^2_{reduced}$ designate the R squared estimates for the more complicated model and the reduced model respectively, N refers to the sample size, and K and $K_{reduced}$ refer to the numbers of predictors for the models (GraphPad Software Inc, 2007; R. Abbott, personal communication, November, 2008).

References

AIMSweb Progress Monitoring and RTI System. (n.d.). Retrieved December 10, 2009, from www.aimsweb.com/measures-2/reading-cbm

Arbuckle, J.L. (1997). *AMOS users' guide version 3.6*. Chicago: SPSS.

Arbuckle, J.L., & Wothke, W. (1995). *AMOS 4.0 user's guide*. Chicago: SmallWaters.

Armbruster, B.B., Lehr, F., & Osborn, J. (2001). *Put reading first: The research building blocks for teaching children to read*. Jessup, MD: National Institute for Literacy.

Chall, J.S. (1996). *Stages of reading development* (2nd ed.). Fort Worth, TX: Harcourt.

Cramer, K., & Rosenfield, S. (2008). Effect of challenge on reading performance. *Reading & Writing Quarterly*, *24*(1), 119–137. doi:10.1080/10573560701501586

Crawford, L., Tindal, G.A., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment*, *7*(4), 303–323. doi:10.1207/S15326977EA0704_04

Daane, M.C., Campbell, J.R., Grigg, W.S., Goodman, M.J., & Oranje, A. (2005). *Fourth-grade students reading aloud: NAEP 2002 special study of oral reading (NCES 2006-469)*. Washington, DC: National Center for Educational Statistics, U.S. Department of Education, Institute of Education Sciences.

Deno, S.L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*(3), 219–232.

Deno, S.L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education*, *37*(3), 184–192. doi:10.1177/00224669030370030801

Deno, S.L., & Marston, D.B. (2006). Curriculum-based measurement of oral reading: An indicator of growth in fluency. In S.J. Samuels & A.E. Farstrup (Eds.), *What research has to say about fluency instruction* (pp. 179–203). Newark, DE: International Reading Association.

Deno, S.L., Mirkin, P.K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children*, *49*(1), 36–45.

Dowhower, S.L. (1987). Effects of repeated reading on second-grade transitional readers' fluency and comprehension. *Reading Research Quarterly*, *22*(4), 389–406. doi:10.2307/747699

Francis, D.J., Santi, K.L., Barr, C., Fletcher, J.M., Varisco, A., & Foorman, B.R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology*, *46*(3), 315–342. doi:10.1016/j.jsp.2007.06.003

Fuchs, L.S., & Deno, S.L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children*, *57*(6), 488–499.

Fuchs, L.S., & Deno, S.L. (1994). Must instructionally useful performance assessment be based in the curriculum? *Exceptional Children*, *61*(1), 15–24.

Fuchs, L.S., Deno, S.L., & Mirkin, P.K. (1984). Effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal*, *21*(2), 449–460.

Fuchs, L.S., Fuchs, D., & Hamlett, C.L. (2007). Using curriculum-based measurement to inform reading instruction. *Reading and Writing, 20*(6), 553–567.

Fuchs, L.S., Fuchs, D., Hosp, M.K., & Jenkins, J.R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, *5*(3), 239–256. doi:10.1207/S1532799XSSR0503_3

Fuchs, L.S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education*, *9*(2), 20–28. doi:10.1177/074193258800900206

Good, R.H., & Jefferson, G. (1998). Contemporary perspectives on curriculum-based measurement validity. In M.R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 61–88). New York: Guilford.

Good, R.H., & Kaminski, R.A. (Eds.). (2002). Dynamic indicators of basic early literacy skills (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement. Retrieved March 2, 2010, from dibels.uoregon.edu

Good, R.H., Kaminski, R.A., & Dill, S. (2002). DIBELS oral reading fluency and retell fluency. In R.H. Good & R.A. Kaminski (Eds.),

Dynamic indicators of basic early literacy skills (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Good, R.H., 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*(3), 257–288. doi:10.1207/S1532799XSSR0503_4

Hasbrouck, J., & Tindal, G.A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher*, *59*(7), 636–644. doi:10.1598/RT.59.7.3

Hintze, J.M., & Christ, T.J. (2004). An examination of variability as a function of passage variance in CBM progress monitoring. *School Psychology Review*, *33*(2), 204–217.

Hintze, J.M., Shapiro, E.S., Conte, K., & Basile, I. (1997). Oral reading fluency and authentic reading material: Criterion validity of the technical features of CBM survey-level assessment. *School Psychology Review*, *26*(4), 535–553.

Hoover, H.D., Dunbar, S.B., & Frisbie, D.A. (2001). *Iowa tests of basic skills*. Itasca, IL: Riverside.

Hosp, M.K., & Fuchs, L.S. (2005). Using CBM as an indicator of decoding, word reading, and comprehension: Do the relations change with grade? *School Psychology Review*, *34*(1), 9–26.

Jenkins, J.R., Hudson, R.F., & Johnson, E.S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review*, *36*(4), 582–600.

Jenkins, J.R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children*, *59*(5), 421–432.

Johnson, E.S., Jenkins, J.R., Petscher, Y., & Catts, H.W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research & Practice*, *24*(4), 174–185. doi:10.1111/j.1540-5826.2009.00291.x

Klauda, S.L., & Guthrie, J.T. (2008). Relationships of three components of reading fluency to reading comprehension. *Journal of Educational Psychology*, *100*(2), 310–321. doi:10.1037/0022-0663.100.2.310

Klein, J.R., & Jimerson, S.R. (2005). Examining ethnic, gender, language, and socioeconomic bias in oral reading fluency scores among Caucasian and Hispanic students. *School Psychology Quarterly*, *20*(1), 23–50. doi:10.1521/scpq.20.1.23.64196

Kranzler, J.H., Brownell, M.T., & Miller, M.D. (1998). The construct validity of curriculum-based measurement of reading: An empirical test of a plausible rival hypothesis. *Journal of School Psychology*, *36*(4), 399–415.

Kranzler, J.H., Miller, M.D., & Jordan, L. (1999). An examination of racial/ethnic and gender bias on curriculum-based measurement of reading. *School Psychology Quarterly*, *14*(3), 327–342. doi:10.1037/h0089012

Kuhn, M.R. (2005). A comparative study of small group fluency instruction. *Reading Psychology*, *26*(2), 127–146. doi:10.1080/02702710590930492

Kuhn, M.R., & Stahl, S.A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology*, *95*(1), 3–21. doi:10.1037/0022-0663.95.1.3

LaBerge, D., & Samuels, S.J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*(2), 293–323. doi:10.1016/0010-0285(74)90015-2

Lesaux, N. (with Koda, K., Siegel, L.S., & Shanahan, T.). (2006). Development of literacy. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the National Literacy Panel on language-minority children and youth* (pp. 75–122). Mahwah, NJ: Erlbaum.

Lipson, M.Y., & Wixson, K.K. (2009). *Assessment and instruction of reading and writing difficulties: An interactive approach*. Boston: Pearson.

Manzo, K.K. (2005). National clout of DIBELS test draws scrutiny. *Education Week*, *25*(5), 1, 12.

Manzo, K.K. (2007). State data show gains in reading. *Education Week*, *26*(34), 1, 27.

Marston, D.B. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M.R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18–78). New York: Guilford.

Marston, D.B., & Magnusson, D. (1988). Curriculum-based assessment: District-level implementation. In J. Graden, J.E. Zins, & M.J. Curtis (Eds.), *Alternative educational delivery systems: Enhancing instructional options for all students* (pp. 137–172). Washington, DC: National Association of School Psychologists.

McGlinchey, M.T., & Hixson, M.D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review*, *33*(2), 193–203.

Mehrens, W.A., & Clarizio, H.F. (1993). Curriculum-based measurement: Conceptual and psychometric considerations. *Psychology in the Schools*, *30*(3), 241–254. doi:10.1002/1520-6807(199307)30:3<241::AID-PITS2310300307>3.0.CO;2-J

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

Newman, H.M. (2009). *The influence of mandated oral reading fluency assessments on teachers' thinking and practice*. Unpublished doctoral dissertation, University of Washington, Seattle.

O'Connor, R.E., & Jenkins, J.R. (1999). Prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading*, *3*(2), 159–197. doi:10.1207/s1532799xssr0302_4

Olson, L. (2007). Instant read on reading, in palms of their hands. *Education Week*, *26*(35), 24, 26, 28, 31, 33–34.

Paris, S.G., Carpenter, R.D., Paris, A.H., & Hamilton, E.E. (2005). Spurious and genuine correlates of children's reading comprehension. In S.G. Paris & S.A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 131–160). Mahwah, N.J.: Erlbaum.

Pearson, P.D. (2006). Foreword. In K.S. Goodman (Ed.), *The truth about DIBELS: What it is, what it does* (pp. v–xix). Portsmouth, NH: Heinemann.

Pearson, P.D., & Johnson, D.D. (1978). *Teaching reading comprehension*. New York: Holt, Reinhart and Winston.

Pedhazur, E.J., & Kerlinger, F.N. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart and Winston.

Pikulski, J.J. (2006). Fluency: A developmental and language perspective. In S.J. Samuels & A.E. Farstrup (Eds.), *What research has to say about fluency instruction* (pp. 70–93). Newark, DE: International Reading Association.

Pikulski, J.J., & Chard, D.J. (2005). Fluency: Bridge between decoding and reading comprehension. *The Reading Teacher*, *58*(6), 510–519. doi:10.1598/RT.58.6.2

Pinnell, G.S., Pikulski, J.J., Wixson, K.K., Campbell, J.R., Gough, P.B., & Beatty, A.S. (1995). *Listening to children read aloud: Oral fluency*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Pressley, M., Hildren, K., & Shankland, R. (2005). *An evaluation of end-grade-3 Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Speed reading without comprehension, predicting little*. East Lansing, MI: Michigan State University, College of Education, Literacy Achievement Research Center.

Rasinski, T.V. (1990). Investigating measures of reading fluency. *Educational Research Quarterly*, *14*(3), 37–44.

Rasinski, T.V. (2006). Reading fluency instruction: Moving beyond accuracy, automaticity, and prosody. *The Reading Teacher*, *59*(7), 704–706. doi:10.1598/RT.59.7.10

Rasinski, T.V., & Hoffman, J.V. (2003). Theory and research into practice: Oral reading in the school literacy curriculum. *Reading Research Quarterly*, *38*(4), 510–522. doi:10.1598/RRQ.38.4.5

Rasinski, T.V., & Padak, N. (2005). *3-minute reading assessments: Word recognition, fluency, & comprehension*. New York: Scholastic.

Rasinski, T.V., & Padak, N. (2008). *From phonics to fluency: Effective teaching of decoding and reading fluency in the elementary school*. Boston: Allyn & Bacon.

Riedel, B.W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly*, *42*(4), 546–567. doi:10.1598/RRQ.42.4.5

Samuels, S.J. (2006). Toward a model of reading fluency. In S.J. Samuels & A.E. Farstrup (Eds.), *What research has to say about fluency instruction* (pp. 24–46). Newark, DE: International Reading Association.

Samuels, S.J. (2007). The DIBELS Test: Is speed of barking at print what we mean by reading fluency? *Reading Research Quarterly*, *42*(4), 563–566.

Schilling, S.G., Carlisle, J.F., Scott, S.E., & Zeng, J. (2007). Are fluency measures accurate predictors of reading achievement? *The Elementary School Journal*, *107*(5), 429–448. doi:10.1086/518622

Schwanenflugel, P.J., Meisinger, E.B., Wisenbacker, J.M., Kuhn, M.R., Strauss, G.P., & Morris, R.D. (2006). Becoming a fluent and automatic reader in the early elementary school years. *Reading Research Quarterly*, *41*(4), 496–522. doi:10.1598/RRQ.41.4.4

Shapiro, E.S. (2000). School psychology from an instructional perspective: Solving big, not little problems. *School Psychology Review*, *29*(4), 560–572.

Sharp, D. (2004). *Supporting teachers' data-driven instructional conversations: An environmental scan of reading first and step literacy assessments, data visualization, and assumptions about conversations that matter: Report to the Information Infrastructure System Project, The John D. and Catherine T. MacArthur Foundation, & the William and Flora Hewlett Foundation*. Retrieved March 3, 2010, from www.dianasharp.com/pdf/report_IISP.pdf

Shinn, M.R., Good, R.H., Knutson, N., Tilly, W.D., & Collins, V.L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review*, *21*(3), 459–479.

Shinn, M.R., Tindal, G.A., & Stein, S. (1988). Curriculum-based measurement and the identification of mildly handicapped students: A research review. *Professional School Psychology*, *3*(1), 69–85. doi:10.1037/h0090531

SPSS. (2004). *SPSS 13.0 base user's guide*. Chicago: Author.

Stage, S.A., & Jacobsen, M.D. (2001). Predicting student success of state-mandated, performance-based state assessment using oral reading fluency. *School Psychology Review*, *30*(3), 407–419.

Valencia, S.W. (in press). Reader profiles and reading disabilities. In R.L. Allington & A. McGill-Franzen (Eds.), *Handbook of reading disabilities research*. Mahwah, NJ: Erlbaum.

Valencia, S.W., Smith, A., Reece, A., Newman, H., Wixson, K.K., & Li, M. (2006, April). *Assessment of oral reading fluency: New models*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Wiley, H.I., & Deno, S.L. (2005). Oral reading and maze measures as predictors of success for English learners on a state standards assessment. *Remedial and Special Education*, *26*(4), 207–214. doi:10.1177/07419325050260040301

Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading*, *5*(3), 211–239. doi:10.1207/S1532799XSSR0503_2

Young, A., & Bowers, P.G. (1995). Individual difference and text difficult determinants of reading fluency and expressiveness. *Journal of Experimental Child Psychology*, *60*(3), 428–454. doi:10.1006/jecp.1995.1048

Yovanoff, P., Duesbery, L., Alonzo, J., & Tindal, G.A. (2005). Grade-level invariance of a theoretical causal structure predicting reading comprehension with vocabulary and oral reading fluency. *Educational Measurement: Issues and Practice*, *24*(3), 4–12. doi:10.1111/j.1745-3992.2005.00014.x

Zutell, J., & Rasinski, T.V. (1991). Training teachers to attend to their students' oral reading fluency. *Theory Into Practice*, *30*(3), 211–217. doi:10.1080/00405849109543502

Submitted August 7, 2009
Final revision received January 31, 2010
Accepted February 22, 2010

Sheila W. Valencia is a professor at the University of Washington, Seattle, USA; e-mail valencia@u.washington.edu.

Antony T. Smith is an assistant professor at the University of Washington, Bothell, USA; e-mail ansmith@uwb.edu.

Anne M. Reece is a graduate of the doctoral program and a teaching associate at the University of Washington, Seattle; e-mail reecea@u.washington.edu.

Min Li is an associate professor at the University of Washington, Seattle; e-mail minli@u.washington.edu.

Karen K. Wixson is a professor at the University of Michigan, Ann Arbor, USA; e-mail kwixson@umich.edu.

Heather Newman is a graduate of the doctoral program at the University of Washington, Seattle; email hnewman00@gmail.com.