

Characteristics of Word Callers: An Investigation of the Accuracy of Teachers' Judgments of Reading Comprehension and Oral Reading Skills

Chad Hamilton
South Lane Public Schools, Cottage Grove, Oregon

Mark R. Shinn
University of Oregon

Abstract. Despite a body of evidence that curriculum-based measurement of reading (R-CBM) is a valid measure of general reading achievement, some school-based professionals remain unconvinced. At the core of their argument is their experience with word callers, students who purportedly can read fluently, but do not understand what they read. No studies have been conducted to determine if teachers' perceptions about these word callers are accurate. This study examined the oral reading and comprehension skills of teacher-identified word callers to test whether they read fluently, but lacked comprehension. Two groups of third graders ($N = 66$) were examined: (a) teacher-identified word callers ($n = 33$) and (b) similarly fluent peers ($n = 33$) who were judged by their teachers to read as fluently as the word caller but who *showed* comprehension. They were compared on R-CBM, CBM-Maze (an oral question-answering test), and the Passage Comprehension subtest of the Woodcock Reading Mastery Test. Results disconfirmed that word callers and their similarly fluent peers read aloud equally well. Word callers read fewer correct words per minute and earned significantly lower scores on the three comprehension measures. Teachers were not accurate in their predictions of either group's actual reading scores on all measures, but were most inaccurate in their prediction of word callers' oral reading scores. Implications for addressing resistance in using CBM as a measure of general reading achievement are discussed.

More than 20 years of research on curriculum-based measurement of reading (R-CBM) has demonstrated that counting the number of words read aloud correctly in 1 minute from standard passages is an excellent measure of general reading proficiency, including reading comprehension. From a traditional psychometric perspective, alter-

nate-form reliabilities typically exceed .90 and 1-week to 1-month test-retest reliability estimates range from .82 to .97 (Good & Jefferson, 1998; Marston, 1989). Criterion-related validity studies typically show correlations of .60 to .80 between R-CBM scores and commercial reading achievement tests and other reading tests (Fuchs, Fuchs,

This research was supported in part by Grant No. 84.029D60057 Leadership Training in Curriculum-Based Measurement and Its Use in a Problem-Solving Model sponsored by the U.S. Department of Education, Office of Special Education Research. The views expressed within this paper are not necessarily those of the USDE.

Address all correspondence and questions about this manuscript to Chad Hamilton, Ph.D., South Lane School District, 455 Adams, Cottage Grove, OR 97424; e-mail: chamilt2@lane.k12.or.us

Copyright 2003 by the National Association of School Psychologists, ISSN 0279-6015

& Maxwell, 1988; Good & Jefferson, 1998; Marston, 1989).

More sophisticated construct validity studies using confirmatory factor analyses have consistently demonstrated that R-CBM scores explain a significant proportion of the variance in reading comprehension construct scores (Petetit, 2000; Shinn, Good, Knutson, Tilly, & Collins, 1992). Additionally, the strong relation of R-CBM as a measure of general reading proficiency has been cross-validated with English language learners (ELL). For example, Baker and Good (1995) reported that correlations between R-CBM and criterion reading measures were comparable for both ELL and English-only students. Similarly, Ramirez (2001) reported that in fifth-grade ELL students, approximately 80% of the variance in reading comprehension construct scores was explained by their English R-CBM reading scores.

Also important, R-CBM has been constructed to satisfy the validity standards from a more contemporary perspective such as the one proposed by Messick (1986). Of these standards, no single standard is more important than that of *consequential validity*; test use should result in decisions that contribute positively to improved outcomes. R-CBM was designed to provide teachers a simple and accurate way of monitoring the progress of their students for purposes of formative evaluation (Deno, 1985, 1986). Repeated studies have shown significant and positive effect sizes in students' achievement when R-CBM is used in formative evaluation (Fuchs & Fuchs, 1986; Lloyd, Forness, & Kavale, 1998). In their meta-analysis, for example, Fuchs and Fuchs (1986) reported effect sizes of .70. This effect size translates into a student who would be expected to be at the 50th percentile when progress is not evaluated formatively to performing at the 76th percentile when this approach is used.

Word Callers as Sources of Teacher Resistance to R-CBM

Despite a well-established body of knowledge establishing R-CBM as a measure of general reading proficiency, including comprehension, some teachers express concern over its use as a general reading

achievement test. This resistance is based on the often-expressed presumption that R-CBM is solely a measure of a student's *decoding* skills. Some teachers express concern cogently when they speak of their experiences with word callers (i.e., students who read fluently, yet do not comprehend what they read). Teachers who have word callers in their classrooms argue that R-CBM may overestimate these students' reading skills and not be sensitive their reading difficulties, in effect mistaking the word callers for competent readers.

The existence and characteristics of word callers has also been a topic in the reading research community (Luftig, 1989). However there are few agreed upon definitions, despite what Stanovich (1986) calls a general acceptance of the concept. Stanovich operationalized word calling as occurring "when the words in the text are efficiently decoded into their spoken forms without comprehension of the passage taking place" (p. 372). Although Stanovich provided a definition, he also expressed skepticism about its validity. He noted that the idea of word callers has gained popularity despite a lack of evidence that applies "to an appreciable number of poor readers" (Stanovich, 1986, p. 372).

The question remains then, are there a group of students (i.e., word callers) whose reading skills would be measured inaccurately by R-CBM? No studies could be found that shed light on this topic. However, given that the observed correlations between R-CBM and other reading comprehension tests are not perfect, it seems plausible that there may be an identifiable subgroup(s) of students for whom the general relation does not apply. Furthermore, it is plausible that one of these subgroups could be students who read fluently, but do not comprehend what they read.

Teacher Judgment of Student Reading Achievement

When teachers raise the topic of word callers, they appear to be basing their objections to R-CBM as a general reading measure based on their judgments of students' reading skills. Therefore, it is essential that these judg-

ments be accurate. It seems intuitive that teachers' impressions that some students read fluently but do not comprehend would be accurate. Teachers base their judgments upon literally "tens of thousands of discrete behavioral events" (Gerber & Semmel, 1985, p. 141). Furthermore, there are empirical studies that support the accuracy of teachers' judgments of student achievement. However, the vast majority of these studies are correlational and have focused on the rank ordering of *groups* of students according to teacher impression and reading tests in one of three forms. First, teachers rated the reading ability for each student in their class, typically on a scale from 1 to 5 (Eaves, Williams, Wincheter, & Darch, 1994; Gresham, Reschly, & Carey, 1987; Hopkins, George, & Williams, 1985; Salvesen & Undheim, 1994; Sharpley & Edgar, 1986; Wright & Wiese, 1988). These ratings then were correlated with students' scores on a reading achievement test. Second, teachers have been asked to predict how their students would perform on a particular measure of reading (e.g., the SRA reading subtests) in terms of grade-equivalent scores (Hoge & Butcher, 1984; Wright & Wiese, 1988), either by rank ordering their students' performance (Brown & Sherbenou, 1981), or by estimating the number of correct answers on a reading test (Helmke & Schrader, 1987). These predictions were then correlated with actual performance. Third, for individual students, teachers surveyed a reading test and predicted which items they believed their students would get correct (Coladarci, 1986; Leinhardt, 1983). The percentage of items for which the teacher was correct was determined.

These types of studies report moderately strong correlations between teacher judgment and the criterion reading measure—ranging from .41 to .86 (median $r = .73$). Other reviews of literature using similar selection criteria have found comparable relations; Hoge and Coladarci (1989) report an average correlation of .67 and Eaves et al. (1994) report a median of .72. However, Hoge and Coladarci (1989) also report distinct variability; teacher accuracy depended upon (a) the teacher, (b) the task judged, and (c) the student judged.

Despite the relatively strong relation between teacher judgment of reading and student achievement on a criterion reading measure, the available research does not provide any evidence of the accuracy of teachers' judgments regarding word callers. Instead, the research has focused on global judgments of reading and rank orderings. Little is disclosed regarding teachers' knowledge of discrete reading skills. That is, the extent to which a teacher can judge a student's oral reading fluency or reading comprehension is unknown.

Purpose of the Study

This study attempts to provide understanding of the reading performance of students that teachers judge as word callers. Two research questions are tested:

1. Are students identified by their teachers as word callers reading fluently but not comprehending?
2. Given that word callers are predicated on teachers' judgments of individual students' oral reading and comprehension skills, are teachers accurate in their judgments of these skills?

Method

Participants

Sixty-six third-grade students and their teachers ($n = 29$) were the participants. Student participants were divided into two groups. The first student group ($n = 33$) was teacher-identified *word callers* (WC). The second student group ($n = 33$) was teacher-identified *similarly fluent peers* (SFP). Participants were from 25 elementary schools from five school districts in the Pacific Northwest. Fifteen of the elementary schools were from a midsize city of over 100,000 residents. Five elementary schools were from smaller communities of approximately 30,000 residents. The remaining five schools were from a small rural community of about 5,000 residents. The university's human subjects committee approved all research practices.

Word callers were identified by their teachers based on a written description that asked "if you teach a third-grade student who can read fluently, but has difficulty comprehending text." The description was based on

the typical language used by teachers to describe their concerns about word callers based on the second author's experience in 20 years of staff development training in CBM. Recruitment began by providing the description on a postcard that was placed in the school mailboxes of 75 third-grade general and special education teachers. Teachers were to return the postcard if they were currently teaching a student who met that description. Thirty-one teachers returned postcards indicating that they had a student who fit the description. Four of the 31 teachers responded that they had two word callers. During a follow-up telephone call, the first author confirmed each teacher nomination by reviewing the reading characteristics of the WC and again asking if the teacher still considered the student a word caller. This process resulted in a confirmation of all 35 potential WC students.

The 33 similarly fluent peers (SFP) then were identified. Each teacher was asked if they also had a student in their classroom who read as fluently as the previously identified word caller, but who had *no* comprehension difficulties. This procedure was conducted independently for *each* teacher-identified word caller to obtain a matched pair. Thus, for the four teachers who had identified two WCs from their classroom, two SFPs were required from the same classroom. Two of the 31 teachers said they did not have similarly fluent student. These two teachers and their two word callers were dropped from the study.

Special education teachers ($n = 4$) who identified a WC were asked to identify a SFP to whom they taught reading. The SFP did not have to be a special education student, but the special education teacher did need to be the student's reading teacher. Three of the special education teachers identified a SFP who received special education; one special education teacher identified a general education student who she instructed in reading. As a result of this process, each of the 33 WC was matched with an SFP from the same classroom, creating a pair of students the teachers believed to have similar reading fluency skills, but different comprehension skills.

The WC group was composed of 27 male and 6 female students (81.8% and 18.2%, re-

spectively). Of the SFP sample, 18 were male and 15 were female (54.5% and 45.5%, respectively). A chi-square analysis confirmed significant gender differences between the groups, $\chi^2(1) = 5.66, p = .017$. The ethnicity of most students was White (64% of the WC and 79% of the SFP group) and both groups were similar in ethnic proportion, $\chi^2(3) = 4.93, p = .177$. Three times as many students were receiving special education services in the WC group than in the SFP group (30.3%, $n = 10$ and 9.1%, $n = 3$, respectively). All special education students received services for less than 50% of the day.

Most teachers (25/29 = 86.2%) taught in the general education classroom; about one-third ($n = 9$) had completed graduate training. The teachers had considerable teaching experience with a reported 13.8 mean number of years, with a range of 2 to 35 years.

Measures

Students were administered four reading tests: (a) Curriculum-Based Measurement of Reading (R-CBM), (b) Curriculum-Based Measurement-Maze (CBM-Maze), (c) a comprehension oral question answering test (CQT), and (d) the Passage Comprehension subtest of the Woodcock Reading Mastery Test (WRMT-PC; Woodcock, 1987). The R-CBM test, CBM-Maze, and CQT test materials were developed from five 400-word traditional folktales used in previous studies of reading comprehension (e.g., Brown & Smiley, 1977; Fuchs, Fuchs, & Maxwell, 1988; Jenkins, Heliotis, Haynes, & Beck, 1986). These folktales had been rewritten by Jenkins et al. (1986) to approximate a second- to third-grade readability level (Fry, 1968) while preserving the theme of the folktale.

Teacher Interview

Before testing, teachers were interviewed to gain information regarding each teacher's quantitative estimates of their students' reading skills on the R-CBM, CBM-Maze, and CQT measures. Each interview lasted approximately 15 minutes. After collecting demographic information, teachers were shown copies of the testing materials that were to be used with their student pairs. The format of the tests was described, including the exact directions. The an-

swers to the CBM-Maze and CQT also were shown and teachers were encouraged to ask questions about the tests and what their students would be expected to do.

To obtain the teachers' quantitative estimates of R-CBM test performance, they were shown the first folktale just as their students would see it, including a cumulative word count. Teachers were asked to point to the place in the passage where the WC would finish in 1 minute. The teacher then was asked how many errors the student would make during that same 1-minute period. These teachers' quantitative judgments were recorded and the process then was repeated for the SFP. For the CBM-Maze, teachers were given the folktale that would be read and they were asked how many items each student would answer correctly in 2 minutes. For the CQT, teachers read the passage and were asked how many questions each student would answer correctly.

Reading Curriculum-Based Measurement (R-CBM)

Three of the five 400-word traditional folktales were randomly selected as reading passages for the R-CBM testing. Standardized R-CBM scoring procedures were used (Shinn, 1989) where students were asked to read aloud from the passage for 1 minute while the examiner scored the number of words read correctly. Words pronounced correctly, given the reading context, and self-corrected errors were scored as correct. Words read incorrectly were those words that were mispronounced, substituted, or omitted. If a student skipped an entire line, each word omitted in that line was scored as incorrect. If students hesitated or struggled with a word for 3 seconds, they were provided the word by the examiner and an error was recorded. The median WRC across three passages was used in the analyses.

Curriculum-Based Measurement-Maze

One folktale from the remaining two folktales was used for the CBM-Maze. The test was constructed using procedures described in Parker, Hasbrouck, and Tindal (1992). The first sentence of the folktale re-

mained intact. Thereafter, every seventh word was replaced by three multiple-choice items consisting of the original story word and two distracters. The two distracters were of the same word length or within one letter of the correct word, but were incorrect with the context of the story. Students had 2 minutes to read the CBM-Maze passage silently and circle correct words. The task was discontinued after the student made three consecutive incorrect choices or the 2-minute time limit had expired. The number of correct maze answers was used in the analysis.

CBM-Maze demonstrates acceptable concurrent validity with commercial reading tests and other reading measures of comprehension. Criterion-related validity correlations with R-CBM are reported from .80 to .89 (Espino, Deno, Maruyama, & Cohen, 1989; Fuchs & Fuchs, 1992, Jenkins & Jewell, 1993). Criterion-related validity with published, norm-referenced tests of comprehension (e.g., Gates-MacGinitie Reading Tests, Stanford Achievement Test-Reading Comprehension subtest) is reported between .77 and .85 (Fuchs & Fuchs, 1992).

Comprehension Oral Question Answering Test (CQT)

Students read the fifth folktale aloud for 3 minutes. As they read, no assistance was given in decoding words, correcting any reading errors, or defining words. If students perseverated on a word for 3 seconds, they were instructed to continue reading. At the end of 3 minutes, students were given the 10 comprehension questions from the Comprehension Oral Questioning Test (CQT) developed by Jenkins et al. (1986). These 10 short-answer comprehension questions were presented orally and students responded orally. According to Jenkins et al. (1986), these 10 questions assessed the *elements of thematic importance* by recall of literal information contained in the folktale. For example, after reading a story about a chief with three sons who challenged them to perform tests of skill, students were asked such questions as "How many sons did the chief have?" and "What did the chief tell his sons to do when they reached the oak tree?" Questions were arranged sequentially by order of occurrence in the folktale. Students' oral responses were transcribed and scored

immediately. Responses were scored as correct if they matched or paraphrased a set of acceptable answers provided by Jenkins et al. (1986). If the student made five consecutive wrong responses, administration was discontinued.

Question answering is the comprehension strategy most commonly employed in classrooms. Despite its frequent use, few technical adequacy studies exist. Fuchs et al. (1988) reported a correlation of .74 between CQT and the Reading Comprehension subtest of the Stanford Achievement Test and correlations with oral and written retell and oral and written cloze from .64 to .76. The correlation with R-CBM was reported to be .84.

Passage Comprehension Subtest of the Woodcock Reading Mastery Tests

The Passage Comprehension subtest (WRMT-PC) from the Woodcock Reading Mastery Tests requires the examinee to read a short passage of two to three sentences silently, and identify orally the key word omitted from the passage. Student responses were scored correct if they matched a list of correct choices identified by the test authors. All participants began with a sample item and continued until six consecutive items were failed or until the last item of the test had been administered.

The test authors report split-half reliability coefficients for the WRMT-PC from .73 to .96 for Grades 3 through 5. Concurrent validity correlations of the WRMT-PC with the Woodcock-Johnson Reading Tests range from .57 to .71 for Grades 3 and 5.

Procedure

Data collectors were the first author and two graduate students in school psychology, all of whom had extensive professional preparation in assessment. Training in administration and scoring of the four reading measures was completed in two sessions, each lasting 1 hour. After training, the two data collectors were observed by the first author and feedback about administration accuracy was provided.

Interscorer agreement was calculated prior to data collection using the overall agreement method (Barlow & Hersen, 1984) where

the number of agreements is divided by the number of agreements + number of disagreements x 100. Interscorer agreement between the first author and the two data collectors was 97% and 98% for R-CBM, 100% and 100% for CBM-Maze, 90% and 100% for the CQT, and 93% and 97% for the WRMT-PC.

During one 20-minute session, all four reading measures were administered individually in the following order: R-CBM passages, R-Maze, CQT, and WRMT-PC. The WC and SFP student pairs were always tested on the same day, although in no predetermined order. Testing typically took place in students' school libraries, and less frequently in a teacher office or quiet area in the back of classrooms.

Results

The means and standard deviations for the four reading measures by group and teacher-predicted score and students' actual score are presented in Table 1. The mean difference (i.e., teacher-predicted score minus student score) also is shown.

A few trends are apparent. The WRMT-PC standard score for the SFP group (99.6) approximates the average score of the national norm sample (48th percentile), lending support that these students did not have reading comprehension problems. In contrast, the mean WRMT-PC score for the WC group was more than one-half standard deviation below the national norm (30th percentile), confirming some difficulty with reading comprehension.

Differences in the CBM-Maze, and CQT comprehension measures also are noted with students in the SFP group outperforming students in the WC group. For interpretative purposes, these differences were converted to effect sizes (ES; Glass & Hopkins, 1996) using the pooled standard deviation. Cooper and Hedges (1994) provide impact criteria to interpret the magnitude of the ES, where .10 is trivial or small, .30 is moderate, and .50 and greater is large. Effect sizes on the three accepted comprehension measures were large, .92 for WRMT-PC, 1.25 for CBM-Maze, and 1.36 for CQT, with SFPs outperforming WCs on all three comprehension measures. A large ES of 1.07 between WCs and SFPs in their R-CBM reading fluency scores also was obtained.

Table 1
Means and Standard Deviations of Differences Between Teacher Estimates and Actual Reading Score by Student Group

Variable	"Word Callers" ^a			Similarly Fluent Peers ^b		
	Teacher Estimate	Actual Score	Mean Difference	Teacher Estimate	Actual Score	Mean Difference
	<i>M</i> (<i>SD</i>)	<i>M</i> (<i>SD</i>)		<i>M</i> (<i>SD</i>)	<i>M</i> (<i>SD</i>)	
R-CBM	129.9 (42.8)	89.1 (19.9)	47.0 (36.4)	138.1 (47.2)	116.2 (22.6)	32.1 (35.4)
CBM-Maze	12.2 (5.0)	9.1 (3.0)	4.8 (3.5)	20.9 (6.2)	13.6 (3.0)	7.5 (5.2)
CQT ^c	5.4 (1.0)	5.2 (1.9)	1.3 (1.1)	8.8 (0.8)	8.2 (1.4)	1.2 (1.3)
WRMT-PC ^d		92.4 (12.7)			99.6 (9.8)	
Accuracy ^e		95.8 (1.6)			97.5 (1.4)	

^a*n* = 33. ^b*n* = 33. ^cScore reflects number correct out of 10 questions. ^dScore reflects standard score, *M* = 100, *SD* = 15. ^eScore reflects accuracy of the oral reading fluency folktale, reported as a percentage.
 Note. R-CBM = Reading Curriculum-Based Measurement; CQT = Comprehension Question Test; WRMT-PC = Woodcock Reading Mastery Test-Passage Comprehension. Standard deviations are in ().

The ES between teacher estimates and student actual scores on CBM-Maze (WC ES = .70 and SFP ES = 1.2) and R-CBM for WC and SFP also was large (WC ES = 1.05 and SFP ES = .57). However, the ES for teacher-predicted CQT score was lower, with a small ES (.13) for the WC group and Moderate ES (.42) for the SFP group. Finally, accuracy rates for both groups were high, exceeding 90%, and similar, 95.8% and 97.5% for WC and SFP, respectively.

Are Word Callers Reading Fluently, But Not Comprehending?

To answer this question, a between-subjects multivariate analysis of variance (MANOVA) comparing the scores of WC and SFP subjects for the R-CBM, CQT, Maze, and WRMT-PC measures was completed. Results showed a significant multivariate main effect, using Pillai's Trace $F(4, 61) = 19.40, p < .001$.

Follow-up univariate analyses indicated significant differences between the two groups on *all four* reading measures, with a consistent pattern of superior performance by the SFP group. The concept of a word caller would be supported by differences on the three commonly accepted measures of reading comprehension (i.e., CBM-Maze, CQT, and WRMT-PC) as observed in this study. However, it would also be expected that there would be *no* differences in R-CBM scores. Based on this analysis, students in the WC group comprehended significantly less well than SFP, but *also* were significantly lower in their oral reading fluency.

Are Teachers Accurate in Judging Students' Reading Skills?

Because teachers predicted student scores on the R-CBM, CBM Maze, and CQT measures, it was possible to examine differ-

ences between these teacher estimates and students' actual scores. Teacher accuracy was tested separately for each measure.

Accuracy of teacher's judgments of oral reading fluency was examined using a two-way, mixed-effects analysis of variance (ANOVA) with one within-subjects effect (type of estimate), and one between-subjects effect (student group). Type of estimate was *teacher prediction* of each student's R-CBM score and *actual* R-CBM score. Cross-validating the previous analysis, a significant difference on R-CBM between the WC and SFP group was observed, $F(1, 64) = 6.60, p < .05$. More important for this question was the observed significant effect of type of estimate, $F(1, 64) = 34.95, p < .01$. Using the Tukey procedure to control for family-wise Type I error, teachers' predicted R-CBM scores were significantly

higher than actual student scores. Results of the student-by-type of estimate interaction for R-CBM also were not significant, $F(1, 64) = 3.22, p > .05$, suggesting that although teachers overestimated R-CBM scores, their overestimation was similar for WC as for SFP. Teachers overpredicted students' reading fluency score regardless of whether the student was judged a WC or a SFP. The differences between mean teacher-predicted R-CBM scores and mean actual R-CBM scores for each group are depicted graphically in Figure 1.

The accuracy of teacher's judgments on the two reading comprehension tests then was examined. For the CQT, another two-way, mixed-effects ANOVA with one within-subjects effect (type of estimate), and one between-subjects effect (student group) was completed. Again, a significant effect of type of estimate

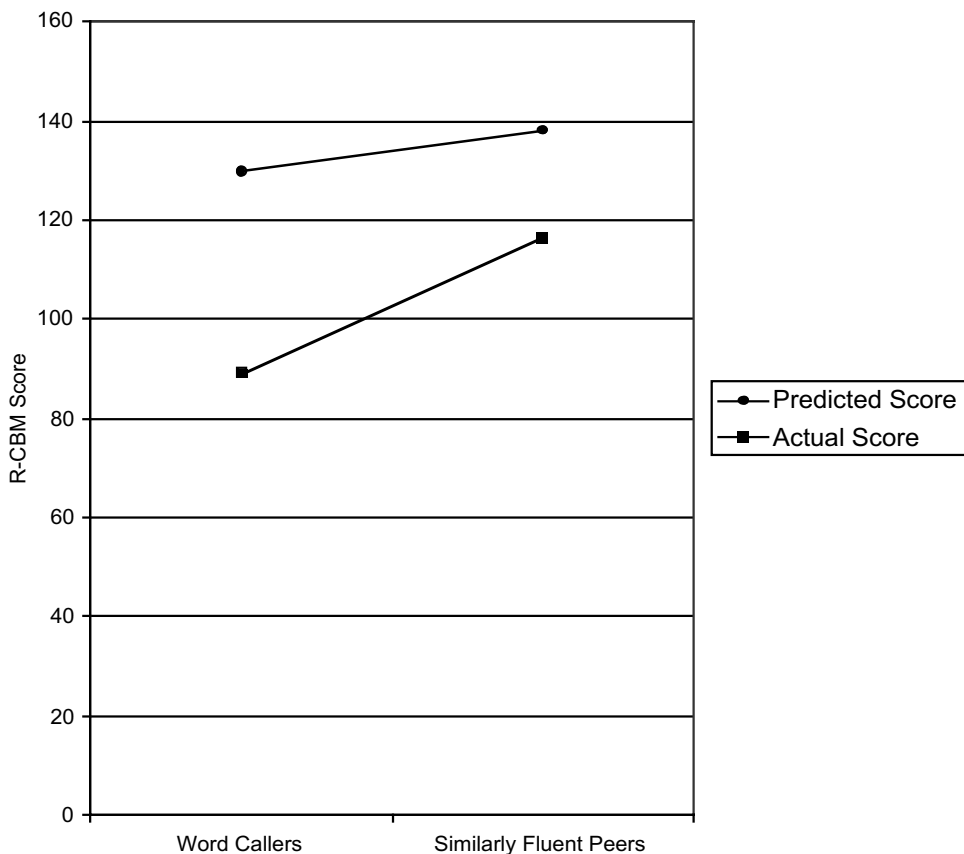


Figure 1. Differences between predicted R-CBM scores and actual R-CBM scores for words callers and similarly fluent peers.

was found, $F(1, 64) = 4.10, p < .05$. Using the Tukey procedure to control for family-wise Type I error, teacher-predicted scores again were significantly higher than actual student scores. Results showed that the student-by-type of estimate interaction for CQT was not significant, $F(1, 64) = 1.26, p > .05$, suggesting that teachers' overestimation of CQT scores was similar for WC and SFP. Differences between mean predicted CQT scores and mean actual CQT scores for each group are depicted in Figure 2. These differences were of a smaller magnitude, especially for the WC group.

For CBM-Maze, another two-way, mixed-effects ANOVA, with one within-subjects effect (type of estimate) and one between-subjects effect (student group), was performed. Using General Linear Models Procedure Least Squares Means, significant differences were found be-

tween predicted and actual scores for both the WC group ($p = .008$) and the SFP group ($p < .0001$). Teachers' predicted scores were higher than actual scores for both groups. However, unlike the analyses with R-CBM and CQT, a significant student-by-type of estimate interaction on the Maze task was observed, $F(1, 64) = 10.17, p < .01$. Teachers significantly overestimated CBM-Maze test scores for students in the SFP group. Differences in predictions, along with differences in actual performance, are depicted in Figure 3.

Discussion

Although a long-standing body of knowledge supports the validity of R-CBM as a general measure of reading achievement, including comprehension, some believe that this measure will not be accurate with a particular type of

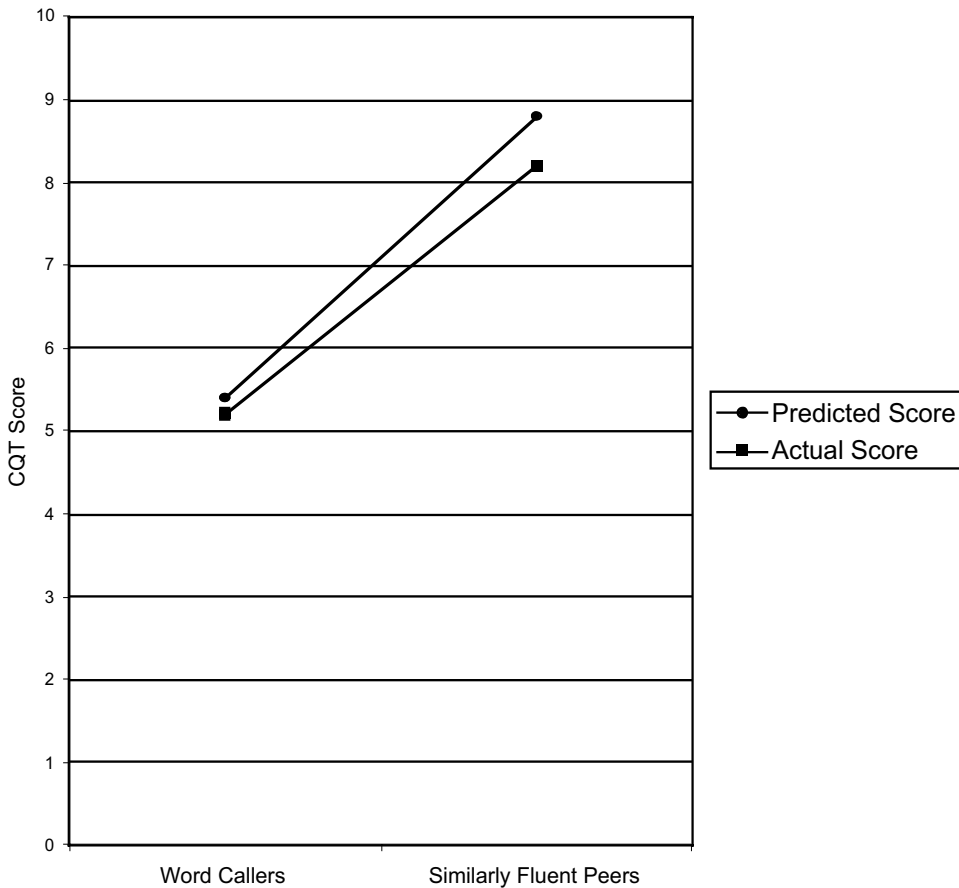


Figure 2. Differences between predicted CQT scores and actual CQT scores for words callers and similarly fluent peers.

student, the word caller. R-CBM results with these students, it is argued, may suggest that they are competent readers when they are not. Thus, there is sometimes resistance to using R-CBM in formative evaluation.

Certainly, acceptance of the existence of word callers, an identifiable group of students who read fluently but do not understand what they read, is not unusual. In this study, 31 third-grade teachers out of 75 (over 40%) identified a student who matched this description. Although this response rate cannot be used as an estimate of national prevalence, it does provide some support for a widespread acceptance of the concept of word callers among teachers. Some discussion of word callers has taken place in the reading literature, although some researchers such as Stanovich (1986, 2000) have expressed skepticism. Furthermore, an examination of the research literature, with and without reference to R-CBM, found no studies on this topic.

This study attempted to examine objectively the reading skills of those students teachers identified as word callers to determine if they read fluently but did not understand what they read. A measure of reading fluency, R-CBM, and three accepted measures of comprehension were given to teacher-identified word callers. Because the basis for teachers' beliefs about word callers was based on their judgments of students' discrete reading skills, it also was important to examine objectively the accuracy of these judgments. A group of students identified by teachers as reading as fluently as WC but who understand what they read, the similarly fluent peers, also was included in this study. This group served the purpose of contextualizing both the reading fluency and comprehension skills of the WC as well as the accuracy of teacher judgments of these skills.

The students identified by their teachers as word callers did *not* fit the expected profile.

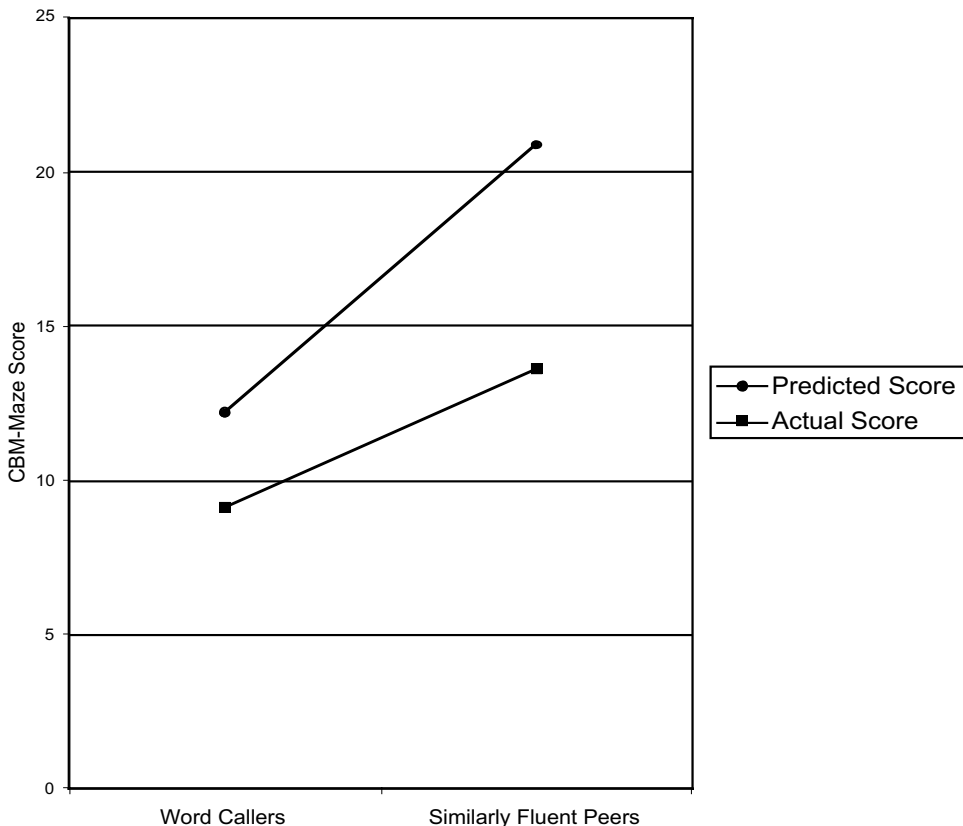


Figure 3. Differences between predicted CBM-Maze scores and actual CBM-Maze scores for words callers and similarly fluent peers.

They did perform more poorly on the reading comprehension measures than students in the SFP group, with large effect size differences. On each of the three measures, the WC group performed almost 1 standard deviation or more lower. However, they also read significantly *less* fluently than students that teachers judged to have equal oral reading skills. This effect size (1.07) with R-CBM also was large.

It appears then, that teachers' judgments about the reading skills of students they consider word callers are half right. The students that teachers judged to be poor comprehenders did perform more poorly than the students teachers identified as having comprehension skills. Additionally, the WC group performed in the very low average range (30th percentile) compared to the normative sample of the Passage Comprehension Test of the Woodcock Reading Mastery Test.

However, even the accuracy of these judgments should be considered suspect. For both WC and SFP groups, teachers overestimated the reading skills on the three measures on which they were asked to provide objective predictions of student performance. Teachers were most accurate in predicting performance on the CQT test. Although their judgments were still statistically significant overestimates, the ES was small to moderate. Whether this relative greater accuracy is attributable to a closer correspondence of the CQT test to classroom practices or is a function of the limited range of the scale (i.e., 1 to 10) should be researched further.

Word Calling As an Idiographic Phenomenon

Because of the lack of exactness of teachers' judgments about reading skills, it is possible that being a WC was a function of teacher holistic impressions; the group design may have masked the idiographic nature of the profile. Objective support for individual word callers, however, was tenuous at best. No student from the WC group read as fluently as the SFP from their classroom, defined as having R-CBM scores within ± 10 WRC. Seven students from the WC group were observed to read at least 10 WRC more than their SFP. Only one of these seven students, however, comprehended more

poorly than their paired SFP. This WC student read 126 WRC, whereas their SFP read 105 WRC. The WC student answered two less questions on the CQT (six vs. eight) than the SFP, identified one less CBM-Maze item, and had one less item correct on the WRMT-PC. This student, of the 33 identified by teachers, *might* be described as fitting the description of a "word caller." However, to say that this student "read fluently, but did not comprehend" would be a liberal interpretation of the definition. The WC student, purported "not to comprehend," was able to answer more than half of the comprehension questions correctly, and scored at the 30th percentile on the WRMT-PC.

In a second attempt to find individual cases of word callers, the comprehension scores of the five students from the teacher-identified WC group who had an R-CBM score at or above the mean rate of SFP group (116 WRC) were examined. These five students might be described as fluent readers. However, these five students also had reading comprehension scores at or above the mean score of the SFP group. It seems that the WC students with higher than predicted oral reading fluency rates also had higher than expected comprehension scores, thus emphasizing the well-documented link between reading fluency and comprehension (Shinn et al., 1992).

What Does Reading Fluency Mean to Teachers?

If the profile of reading fluently but not comprehending was not evident in the students teachers identified as word callers in this study, then what can account for the teachers' perceptions? The general overestimation of students' scores on the four reading measures does not explain the overall accurate teachers' judgments that students in the WC group comprehended less well than their similarly fluent classmates. Perhaps teachers confuse accuracy with fluency or they confuse reading any words aloud correctly with fluency. With respect to the former, on the R-CBM third-grade passages, the WC group had a mean accuracy rate of 95.8%, and the SFP group read with 97.5% accuracy. These rates are high and similar. Perhaps, teachers are using the terms accuracy

and fluency interchangeably. With respect to the latter, students in the WC group read an average of 89.1 words correctly per minute. This level of performance is clearly different than reading 0 words correctly.

Either explanation seems plausible. It is possible that during reading instruction and incidental observation, teachers hear some students reading words correctly and/or read words with few errors. The teachers may conclude that these students are “fluent,” yet these same students are dissimilar from their peers in terms of comprehension skills. More research should be conducted to explore teachers’ meaning of reading fluency to gain more understanding about what teachers understand and use to form their judgments.

Implications and Limitations

As is the case with any research study, the conclusions drawn must be tempered by the study’s limitations. No single piece of research provides definitive answers and must be replicated. Foremost of the limitations is external validity. Participants were third-grade students and their teachers from the Northwest United States. The generalizability to other geographic areas, grades, and students should be investigated further. Additionally, although the number of standardized measures of reading fluency is limited in number, there are many other methods of assessing reading comprehension than the ones employed in this study. This study should also be replicated using different measures of reading comprehension.

The results of this study potentially hold important practical implications for school-based professionals. Those who find it difficult to accept R-CBM as an index of general reading achievement often cite their own experiences regarding a student who “reads fluently, but can’t comprehend.” This study demonstrated there were differences in comprehension and oral reading fluency between students identified by teachers as word callers and students who had comprehension skills. Furthermore, teachers significantly overpredicted the oral reading skills of students in both groups. It seems that something else is going on with respect to teachers’ perceptions of word callers and oral reading flu-

ency, but the impression that the students are reading fluently is not an accurate one. As Gerber and Semmel (1985) suggest, teachers may observe thousands of student behaviors per day. It seems likely, however, that teachers may be better at rank ordering students by general reading skills than providing accurate and specific judgments about students’ discrete reading skills. Rather than simply accepting these judgments, it seems logical for practitioners to emphasize the need for objective measures. In the case of oral reading fluency, teachers’ impressions can be cross-validated in less than 5 minutes, and notably, each of the more explicit comprehension measures can be completed in less than 10 minutes total.

References

- Baker, S. K., & Good, R. (1995). Curriculum-based measurement of English reading with bilingual Hispanic students: A validation study with second-grade students. *School Psychology Review, 24*(4), 561-578.
- Barlow, D. H., & Hersen, M. (1984). *Single case experimental designs* (2nd ed.). Boston: Allyn & Bacon.
- Brown, L. L., & Sherbenou, R. J. (1981). A comparison of teacher perceptions of student reading ability, reading performance, and classroom behavior. *The Reading Teacher, 34*, 557-560.
- Brown, A. L., & Smiley, S. S. (1977). Rating the importance of structural units of prose passages: A problem of meta-cognitive development. *Child Development, 48*, 219-232.
- Coladarci, T. (1986). Accuracy of teacher judgments of student responses to standardized test items. *Journal of Educational Psychology, 78*, 141-146.
- Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219-232.
- Deno, S. L. (1986). Formative evaluation of individual student programs: A new role for school psychologists. *School Psychology Review, 15*, 358-374.
- Eaves, R. C., Williams, P., Wincheter, K., & Darch, C. (1994). Using teacher judgment and IQ to estimate reading and mathematics achievement in a remedial-reading program. *Psychology in the Schools, 31*(3), 261-272.
- Espin, C., Deno, S. L., Maruyama, G., & Cohen, C. (1989, March). *The Basic Academic Skills Samples (BASS): An instrument for the screening and identification of children at risk for failure in regular education classrooms*. Paper presented at the annual conference of the American Educational Research Association, San Francisco.
- Fry, E. (1968). A readability formula that saves time. *Journal of Reading, 11*, 513-516, 575-578.
- Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*, 199-208.

- Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21*, 45-58.
- Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*(2), 20-27.
- Gerber, M. M., & Semmel, M. I. (1985). Teacher as imperfect test: Reconceptualizing the referral process. *Educational Psychologist, 19*, 137-148.
- Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology*. Boston: Allyn & Bacon.
- 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: The Guilford Press.
- Gresham, F. M., Reschly, D. J., & Carey, M. P. (1987). Teachers as "tests": Classification accuracy and concurrent validation in the identification of learning disabled children. *School Psychology Review, 16*, 543-553.
- Helmke, A., & Schrader, F. W. (1987). Interactional effects of instructional quality and teacher judgment accuracy on achievement. *Teaching and Teacher Education, 3*, 91-98.
- Hoge, R. D., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement levels. *Journal of Educational Psychology, 76*, 777-781.
- Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*, 297-313.
- Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement, 22*, 177-182.
- Jenkins, J. R., Heliotis, J., Haynes, M., & Beck, K. (1986). Does passive learning account for disabled readers' comprehension deficits in ordinary reading situations? *Learning Disability Quarterly, 9*, 69-76.
- Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children, 59*, 421-432.
- Leinhardt, G. (1983). Novice and expert judgment of individual student achievement. *Educational Psychologist, 18*, 165-179.
- Lloyd, J. W., Forness, S. R., & Kavale, K. A. (1998). Some methods are more effective than others. *Intervention in School and Clinic, 33*, 195-200.
- Luftig, R. L. (1989). *Assessment of learners with special needs*. Boston: Allyn & Brown.
- Marston, D. B. (1989). A curriculum-based measurement approach to assessing academic performance: What to do and why to do it. In M. R. Shinn (Ed.), *Curriculum-based measurement* (pp. 18-78). New York: The Guilford Press.
- Messick, S. (1986). *The once and future issues of validity: Assessing the meaning and consequences of measurement* (Research Rep.). Princeton, NJ: Educational Testing Service.
- Parker, R., Hasbrouck, J. E., & Tindal, G. (1992). The maze as a classroom-based reading measure: Construction methods, reliability, and validity. *The Journal of Special Education, 26*, 195-218.
- Petitet, L. (2000). *Construct validity of curriculum-based measurement of reading: Contributions of method variance*. Unpublished doctoral dissertation. University of Oregon, Eugene.
- Ramirez, C. R. (2001). *An investigation of English language and reading skills on reading comprehension for Spanish-speaking English language learners*. Unpublished doctoral dissertation. University of Oregon, Eugene.
- Salvesen, K. A., & Undheim, J. O. (1994). Screening for learning disabilities with teacher rating scales. *Journal of Learning Disabilities, 27*, 60-66.
- Sharpley, C. F., & Edgar, E. (1986). Teachers' ratings vs. standardized tests: An empirical investigation of agreement between two indices of achievement. *Psychology in the Schools, 23*, 106-111.
- Shinn, M. R. (1989). *Curriculum-based measurement*. New York: The Guilford Press.
- Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21*, 459-479.
- Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360-407.
- Stanovich, K. E. (2000). Romance and reality. In K. E. Stanovich (Ed.), *Progress in understanding reading* (pp. 392-404). New York: The Guilford Press.
- Woodcock, R. M. (1987). *Woodcock Reading Mastery Tests-Revised*. Circle Pines, MN: American Guidance Corp.
- Wright, D., & Wiese, M. J. (1988). Teacher judgment in student evaluation: A comparison of grading methods. *Journal of Educational Research, 82*, 10-14.

Chad Hamilton, Ph.D., is a school psychologist in the South Lane School District, Cottage Grove, Oregon. He received his Ph.D. from the University of Oregon in 2001. His research interests include formative evaluation, Curriculum-Based Measurement (CBM), and how teacher judgments influence educational assessment.

Mark R. Shinn, Ph.D., is Professor in the Special Education Area at the University of Oregon. He received his Ph.D. from the University of Minnesota in 1981. His teaching and research interests include educational assessment and Curriculum-Based Measurement (CBM) and its use in a problem-solving model. He consults regularly with school systems around the country on problem-solving service delivery systems and effective problem-solving teams. More information can be found at <http://darkwing.uoregon.edu/~mshinn/>