2012

Practical Issues in Field Based Testing of Oral Reading Fluency at Upper Elementary Grades

Luke Duesbery

Jenelle Braun-Monegan Stone

Jacob Werblow

Drew Braun

Practical issues in field based testing of oral reading fluency at upper elementary grades

Luke Duesbery [a], Jenelle Braun-Monegan [b], Jacob Werblow [c,*], Drew Braun [d]

[a] San Diego State University, United States
[b] Walden University, United States
[c] Central Connecticut State University, United States
[d] Bethel School District, United States

ABSTRACT

In this series of studies, we explore the ideal frequency, duration, and relative effectiveness of measuring oral reading fluency. In study one, a sample of 389 fifth graders read out loud for 1 min and then took a traditional state-level standardized reading test. Results suggest administering three passages and using the median yields the highest predictive validity. Study two compared oral reading fluency rates at 30 and 60 s for 815 elementary and middle school students on the same passage. Results indicate that the 30 s measures yield a comparable score. Study three found relatively similar predictive validity of oral reading fluency for 67 fourth- and 125 sixth-grade students on Aimsweb, EasyCBM, and DIBELS. Implications for practice are discussed.

Keywords:
Oral reading fluency
Curriculum-based measurement
Predictive validity
Efficiency

Introduction

Curriculum-based measures (CBMs) emerged from the special education model of using repeated measurement data to monitor progress and inform instruction (Deno, 2003), but over the last 30 years the purpose and use of CBMs have expanded into regular education classrooms and beyond. CBMs are commonly used by classroom teachers to monitor the progress of all students, inform instruction, and make adjustments in placements and programs throughout the school year. At the district level, CBMs may be used by administrators to evaluate programs, predict success on high-stakes tests, screen students who are at risk, monitor students' progress, and establish criterion levels of behavior (Crawford, Gerald, & Stieber, 2001; Deno, 2003; Helwig, Heath, & Tindal, 2000; Stage & Jacobsen, 2001). Thus, in a large number of school districts around the United States, CBMs have become an integral part of instructional programs. With leading scholars (e.g., Fuchs and Fuchs) and the U.S. Department of Education (e.g., IDEA) propagating instructional decision-making models such as Response to Intervention (RTI) and Scientifically Research Based Interventions (SRBI), CBMs will likely continue to gain popularity in the field.

Early in their adoption, Marston (1989) postulated that CBMs should possess a number of critical attributes: The measures should be quick to administer, directly observe student behavior, have potential for equivalent forms, be inexpensive to administer, and show sensitivity to growth over time. Marston advocated for *efficiency* and *effectiveness*. More recently, however, scholars have focused on the importance of these elements in relative isolation. For example, MacMillan and Fewster (2002) argued that the features that make CBMs effective are sensitivity to change, ease of administration, and ease of creating equivalent forms (efficiency). Fuchs, Fuchs, Hosp, and Hamlett (2003) delineated standardization, long range monitoring, and a focus on grade-level reading as the integral components of CBMs (effectiveness).

Reschly, Busch, Betts, Deno, and Long (2009) meta-analysis of over 30 years of research further validated that CBMs have a significant, strong correlation ($r = 0.67$) with student performance on other standardized reading tests; however, they also concluded that this relationship is potentially moderated by other variables, including the number of reading passages given, sources of the passages (different testing companies), and the psychometric properties of different passage sets. Thus there is a need to determine to what extent the relationship between CBMs and other standardized reading tests is confounded by other factors.

Furthermore, evidence suggests that CBMs in reading may not be appropriate measures for progress monitoring due to some severe psychometric limitations. In an analysis of popular CBM testing companies used for progress monitoring, Ardoin and Christ (2009) found that a 68% confidence interval around a 1.5 words per minute observed rate of growth indicates that the individual students' actual growth could be anywhere from excellent

(1.36 + 1.08) to inadequate (1.36 – 1.08). These findings put into question whether or not reliable decisions can be made regarding individual students' response to instruction. The authors emphasized the importance of using multiple ORF passages per session to reduce error.

Oral reading fluency

In this series of studies, we explore some of these potential moderating factors and also return to the critical notions in CBM (as in the early work of Marston) by addressing both CBM efficiency and effectiveness. Given that CBMs are now being implemented in numerous school districts system-wide and in multiple subject areas across the United States, they require a larger commitment of resources that may result in lost instructional time; hence efficiency and effectiveness become more important. In this series of three studies, we focus on the most commonly used CBM in reading: oral reading fluency (ORF).

Published ORF measures are individually administered and consist of asking students to read a standardized grade-level passage out loud for 1 min. As the student reads, the administrator tracks errors, or incorrect words read, and marks the final number of words read correctly when the 1 min time limit ends. Self-corrected words within 3 s are counted as accurate, while hesitations longer than 3 s result in a word error. The total number of words read minus the number of errors equals the score for the 1 min test (Children's Educational Services, 1987; Hosp, Hosp, & Howell, 2007; Shinn & Bamonto, 1998). Although this is the general administration procedure for ORF passages, each published assessment system varies slightly in prescribed administration procedure.

Psychometric properties. Oral reading fluency is known to be an effective indicator of reading skill (Fuchs, Fuchs, Hosp, & Jenkins, 2001), and the psychometric properties of ORF probes have been strengthened and supported over time by a number of scholars in the field (see Alonzo & Tindal, 2009; Jenkins & Jewell, 1993; Reschly et al., 2009; Tindal & Marston, 1996). Criterion levels of validity have been found to be in the range of $r = 0.73–0.91$ (Deno, Mirkin, & Chiang, 1982). Alternate forms reliability at the middle school level has been documented in the range of 0.88–0.92 (Barth et al., 2012). ORFs are widely considered to be an empirically sound measure of students' reading skill; however, there are many potential uses of ORF data, and each usage should be validated for informing decisions within its appropriate context. For example, Christ (2006) examined the stability of progress monitoring outcomes and CBM-R slopes at the level of the individual student and found sizeable estimates of the standard error of slopes (0.78 words).

Purpose. This paper seeks to build from the recommendations from Reschly et al.'s (2009) meta-analysis by exploring the potential moderating variable of time and number of probes used during administration, and address other practical questions regarding the use of ORF as a school-wide screening or benchmarking tool though the examination of three related studies.

Research questions

From a practical standpoint, we address the issues of efficiency and effectiveness with three empirical studies. *Study one* explores efficiency by asking if schools can save time and/or money by reducing the number of ORF probes administered without reducing the quality of data collected. Schools that screen all students using ORF may save substantial time and money if the standard practice of administering three 1 min ORF passages can be reduced to administering one or two 1 min passages without compromising validity and reliability. In *study two*, we again address efficiency as we explore optimal ORF duration and ask if schools can save time or money by abbreviating the measure. If the standard 1 min ORF probe is reduced to 30 s without compromising student placement accuracy, schools could further save time and money. In *study three*, we explore the relative effectiveness of ORF measures from three leading CBM assessment systems (Aimsweb, EasyCBM, and DIBELS). We ask which measure of ORF is the most reliable and valid for predicting end of the year state assessment performance.

Because this research took place in the course of field-based practice, these studies were conducted on different samples and at different times over a 3-year period. Notable differences are described.

Study one: efficiency through changes in frequency

In an era of increasing school accountability, instructional time has become more precious due to the time now spent preparing for and administering standardized tests. Although the cost associated with implementing a district-wide assessment program is complex (involving planning, resources, and staff/parent volunteers), this study concentrates only on the cost incurred as a result of lost instructional time due to multiple test probes used during administration. Currently, as indicated by the administrative directions of Aimsweb and DIBELS, the standard in the field is to administer three 1 min ORF passages to each student and to use the median score to identify candidates for instructional intervention. The rationale for this approach is likely rooted in classical test theory: more items lead to lower measurement error and higher test reliability; however, ORF administration would be far more cost efficient if only one or two passages could arrive at similar results, without loss in measurement precision. For example, Ardoin et al. (2004) concluded that using a single probe was sufficient for the purposes of universal screening.

Method

Participants

The district in which this study took place, a medium size school district ($n = 5600$) in the Pacific Northwest of the United States, administers ORF probes to all students in grades one through eight in the fall, winter, and spring. All students, including students with mild disabilities and English language learners, participate in the district ORF assessment. Three ORFs were administered to each of 389 students in the fifth grade. In our sample, approximately 77% of students were Caucasian, 14% Hispanic, 3% Asian/Pacific Islander, 3% American Indian/Native Alaskan, 2% African American, and 1% other. Fifty-five percent of these students met federal guidelines for free and reduced-price meals, and 14% were enrolled in special education. There were approximately equal numbers of males and females. This sample size was sufficient for the statistical analyses described below.

Fifth grade was chosen because of the opportunity to use the state large-scale reading test, administered at the end of the year, as a stable criterion reference. The state reading test is psychometrically sophisticated as it was developed using Item Response Theory (IRT) and has well documented technical adequacy (see Oregon Department of Education, 2007).

Procedure

Students were administered grade-level appropriate ORF passages selected from the Test of Oral Reading Fluency (TORF) (Marston & Deno, 1987). In this district, the TORF passages have been used for 10 years. District administrators explained that the primary reason for selecting passages from the TORF was to ensure the construct did not drift over time, and difficulty levels remained constant (Fuchs & Fuchs, 2002). It is imperative that alternate passages administered throughout the year were equivalent so that "changes in student scores are attributable to student improvement rather than changes in testing conditions" (Shinn & Bamonto, 1998, p. 6). According to district administrators, teachers were confident the passages accurately measured student reading rates and lead to appropriate placement decisions. All district-employed test administrators were required to attend a 3 h district-wide training in standardized district ORF protocols. Across fifth grade, the same passages were administered in the same order.

In this district, instructional reading intervention categories were based on a series of ORF cut scores derived both from district data and norms from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) (Good & Kaminski, 1996). Cut scores were also refined based on prior student performance. *Established* readers had a greater than 90% probability of passing the end-of-year state reading assessment. In contrast, *intensive* readers had a less than 10% probability of passing. These low-performing fifth grade students, reading fewer than 70 words per minute, were typically eligible for an alternative reading program in lieu of the regular district reading program. Actual student placement was determined by triangulating results from this ORF screen, a curriculum test, and teacher recommendation.

To answer our research question, "How many ORF probes are necessary?" we started with the assumption that the standard by which we make all comparisons should be the current state of affairs. Currently, the district administers three separate 1 min ORF measures to each student in the fall and makes student placement decisions based on the median value. This may end up being the first, second, or third ORF probe. We will refer to the current practice in the district as method one; we will then compare it to four alternative methods outlined in Table 1.

Method two employs a single ORF probe. For method three, using a larger dataset obtained the prior school year (adjusted $r^2 = 0.9$, $p < 0.001$), we regressed the spring scores obtained by the traditional method of administering three probes and taking the mean onto the scores obtained from administering only the first probe. The equation for the predicted score was $y = 0.996 \times$ (first probe result) + 5.34. This equation might be simplified in the field to: ORF = (first probe) + 5. Method four uses the mean of two ORF probes, and method five uses the higher of two ORF probes.

We evaluated the four alternate methods by comparing the placement decisions made for students under the original method with commensurate placement decisions resulting from the alternates. For example, in method one, students were placed into one of three intervention categories based on the median of three ORF measures. In the four alternate methods, we compared the resulting placement decisions to method one, and evaluated the impact of the change in placement on the student population. A sensitivity and specificity analysis was used to address this question. Sensitivity indicates the proportion of students identified for intervention who did not meet the state benchmark: *true positives*. Specificity, on the other hand, indicates the proportion of students who are not identified for intervention who met the state benchmark: *true negatives*.

Table 1
Alternate methods of measuring oral reading fluency.

	ORF calculation method
Method 1	Median of three ORFs (current practice)
Method 2	One ORF
Method 3	Prediction from regression based on one ORF
Method 4	Mean of two ORFs
Method 5	Highest of two ORFs

Table 2
5th grade students in each category by method with sensitivity and specificity.

	% Low	% Mid	% High	Sensitivity	Specificity
Method 1	8.7	59.9	31.4	91.4%	55.6%
Method 2	4.6	57.8	37.5	81.0%	58.6%
Method 3	3.3	56.3	40.4	79.3%	63.1%
Method 4	2.1	33.4	64.5	86.2%	61.1%
Method 5	4.4	56.8	38.8	81.0%	63.1%

Results and discussion

Across individual tests and administrations, test–retest reliabilities ranged from $r = 0.92$ to 0.96, and internal consistency reliabilities were around $r = 0.98$. Table 2 presents the hypothetical proportion of students assigned to each intervention category by each method and the sensitivity and specificity results for the fall administration period.

The sensitivity analysis is the most important metric for identifying students at risk of not meeting the end of the year benchmark. Based on these results, the district's current method is the best method for screening students who might be in danger of not meeting the benchmark (measured by the end of the year state test). Administering three ORF probes to each student and selecting the median score was the most sensitive decision-making procedure, accurately identifying about 91 in every 100 students. Method four (the mean of two probes) is the next most sensitive procedure, identifying about 86 in every 100 students accurately. In terms of the potential for reduced administration time, methods two and three were the most promising. Both methods require only a single ORF probe during administration; however, they are less accurate, identifying about 79 in every 100 students.

The identification accuracy associated with the number of ORF probes is a matter that needs further discussion. Ideally, the most accurate method should be standard practice, but the increased accuracy associated with three ORF probes may not be worth the cost. Adopting method two would save having to administer a third passage to each student, saving hundreds of hours of instructional time in even a small school district over the course of the year. In doing so, however, we would fail to identify a small proportion of students who are in danger of not meeting the end of year benchmark. In our current method, about 9 in 100 students are not identified. By moving to method two, the ratio would increase to 14 in every 100 students. This evidence clearly points to the potential for an appreciable gain in time savings, but this savings is coupled with a loss in identification accuracy.

Study two: efficiency through changes in duration

A 1 min ORF probe is standard in the field. This duration was probably chosen because 'one-minute' is easy to remember (G. Tindal, personal communication, February, 2008); however, at least one study has suggested that a 30 s administration may be comparable (Fuchs, Tindal, & Deno, 1981). In this abovementioned study ($n = 45$), correlations between 30 and 60 s ORFs ranged between $r = 0.92$ and 0.96 (page 24), suggesting these two measures might be interchangeable. Additionally, there was evidence to suggest the 30 s ORF might be preferable given decreased variability over multiple passages with the shorter measure. The authors, however, noted concerns over the counter-intuitive finding that increasing the ORF to 3 min further reduced variability. In the present study, we return to the topic of ORF duration because innovations and improvements in test form comparability, coupled with a more robust sample size, may provide a more clear answer.

Method

Participants

Our sample included 815 students from grades two, three, four, five, seven, and eight from a medium size rural school district in the Pacific Northwest of the United States. Participants consisted of approximately equal numbers of males and females. All students in these schools, including students with mild disabilities, were included in analyses. The sample demographic makeup was approximately 81% Caucasian, 11% Hispanic, 2% Asian/Pacific Islander, 3% American Indian/Native Alaskan, 2% African American, and 1% other. About 60% of students met federal guidelines for free and reduced price meals, and 15% were enrolled in special education.

Procedure

Each participant was administered a single standard grade-level ORF probe from the Test of Oral Reading Fluency (Marston & Deno, 1987). The test administrator marked the number of words read at both 30 and 60 s on the test protocol. To examine the degree of consistency between the two measures, we employed both a relatively conservative intraclass correlation coefficient (ICC) (two way-random effects consistency type model (Shrout & Fleiss, 1979), and a two-tailed Pearson correlation). All analyses were modeled with SPSS Version 20.0.

Results and discussion

Average words read per minute varied over the entire sample of 815 students, with a low at second grade of 119 (SD = 41) and a high at seventh grade of 167 (SD = 39). At each grade, the mean 30 s score was a little more than 50% of the mean 60 s score. More specifically, at second grade it was 53%, third grade 53%, fourth grade 54%, fifth grade 51%, seventh grade 54%, and at eighth grade 54%. Between the 30 and 60 s measures, the degree of consistency modeled with the ICC was high, ranging from $r = 0.846$, $p < 0.001$ at seventh grade, to $r = 0.891$, $p < 0.001$ at eighth grade. Using the more traditional Pearson correlation, the degree of consistency was very high, ranging from $r = 0.91$, $p < 0.001$ at seventh grade, to a near perfect correlation of $r = 0.971$, $p < 0.001$ at fourth grade. These results closely parallel the correlations of sixth graders ($r = 0.92$) found in the Fuchs et al. (1981) study. Both the conservative and traditional correlations were high, pointing to comparability of the differently timed fluency measures. A full summary of results is provided in Table 3.

Study three: relative effectiveness

As a result of the heightened awareness of CBMs, many organizations and researchers are developing and marketing such measures for teachers and school district personnel. Test publishers such as Pearson and CTB McGraw-Hill have developed CBMs in key academic domains in association with various research organizations funded through federal grants and/or private support. Although some key features of CBMs are common, each organization proceeds with development using different sampling plans, timing and administration procedures, and measurement models. These differences may impact the comparability of scores and, therefore, the interpretation of student learning and growth. As the importance of CBM as a measurement tool has grown and the stakes in the decisions being made with these instruments mount, careful review of the comparability of common measures across test publishers is warranted.

Table 3
Comparison of 30 and 60 s probes of oral reading fluency across grades.

Grade	Age range	*n*	30 s mean (SD)	60 s mean (SD)	Pearson *r*	ICC
2	7–8	123	63 (21)	119 (41)	0.970*	0.879*
3	8–9	128	68 (21)	128 (40)	0.951*	0.877*
4	9–10	128	81 (23)	151 (44)	0.971*	0.884*
5	10–11	129	78 (25)	153 (46)	0.945*	0.883*
7	12–13	142	90 (21)	167 (42)	0.910*	0.846*
8	13–14	165	88 (22)	163 (39)	0.931*	0.891*

* $p < 0.001$.

Normative rates typically serve as the referent for making screening and progress monitoring decisions using ORF, as in the benchmarks used by the district in study one. Although some publishers, such as Pearson, create their own normative referents, others use published norms, such as those established by Hasbrouck and Tindal (1992, 2006). Furthermore, because of the inconsistency across CBM norms, score comparison becomes questionable. For this reason, this study aims to examine the comparability of three widely used ORF assessment systems. Specifically, we investigate the following research question about the relative effectiveness of Aimsweb, EasyCBM, and DIBELS: are similar grade-level passages of comparable difficulty, and do ORF passages predict reading comprehension with comparable accuracy?

Method

Participants

We administered progress monitoring ORF probes from each of three assessment systems (Aimsweb, EasyCBM, and DIBELS) to two samples of students: 125 sixth-grade students from the same high poverty urban school used in study two and 67 fourth-grade students from a relatively affluent school district in Southern California, both in the United States. The sample from the more affluent school district was 73% Caucasian, 14% Asian, 5% Hispanic or Latino, and 4% African American. We used the high poverty sample to answer the question about the relative difficulty of ORF passages and the more affluent sample of students to answer the question about relative predictive validity.

Procedure

Our study compared three measurement systems: Aimsweb, published by Pearson; EasyCBM, created by Behavioral Research and Teaching at the University of Oregon and published by Riverside; and DIBELS, distributed by the University of Oregon and Sopris West. Aimsweb has approximately 30 equivalent forms of ORF probes at each grade level. According to the publisher, experienced educators wrote the passages that were subsequently field-tested and revised. Readability formulae were used to establish grade-level criteria. For the grades used in this study, alternate form reliability ranged from $r = 0.84$ to 0.85 (Howe & Shinn, 2002). EasyCBM has approximately 20 equivalent forms at each grade level. For the grades used in this study, alternate form reliability ranged from $r = 0.92$ to 0.94 (Alonzo & Tindal, 2009). Passages were written by graduate students and were reviewed and field-tested. Readability at each grade level is targeted for mid-year, thus a 6th grade passage should have a readability of 6.5. DIBELS has approximately 20 alternate forms of the ORF at each grade level. For lower grades, alternate form reliability was reported at $r = 0.94$ (Good, Kaminski, Smith, & Bratten, 2001), although we could not find reliability coefficients for the 4th grade passages we used in this study. For the three alternate forms used from each assessment system in this study, average inter-correlations were $r = 0.88$ for Aimsweb, $r = 0.83$ for EasyCBM, and $r = 0.86$ for DIBELS. All correlations were significant ($p < 0.01$).

Table 4
Internal consistency within ORF family for 6th grade ($n = 125$, high poverty sample).

	Duration (s)	Mean words per minute	SD	ICC	Pearson r
Aimsweb	30	46.5	15.4	0.73*	0.94*
	60	91.3	32.1		
EasyCBM	30	47.6	15.4	0.77*	0.94*
	60	95.6	29.5		
DIBELS	30	44.8	15.5	0.77*	0.93*
	60	86.5	29.1		

* $p < 0.001$.

We chose the California State Test of Language Arts as our dependent measure for regression analysis. Internal-consistency reliability for the test was $r = 0.94$ (California Department of Education, 2009). At the fourth grade level, approximately 24% of test items target word analysis, fluency, and vocabulary knowledge. Twenty percent target skill in reading comprehension, 12% cover literary analysis skill, and the remaining address writing conventions (California Department of Education, 2002).

Given the results of study two, we elected to administer nine 30 s ORFs to each student: three from Aimsweb, three from EasyCBM, and three from DIBELS. All students were given the grade-level passages in one administration period. To prevent contamination by different administration procedure, a single administration procedure was used across the three ORF assessment systems. The order of administration of each ORF family was counterbalanced to counteract any order effect. Two school psychologists, highly experienced administering assessments, administered the passages. The relative predictive nature of the ORFs was compared with linear regression, regressing the median of three 30 s ORF scores onto the end of year state reading test.

Results and discussion

In terms of consistency within the assessment systems, we discovered that all three were of high quality. Results presented in Table 4 suggests that the measures from each assessment system had high internal consistency reliability, all above $r = 0.93$. In terms of relative difficulty, over a sample of three passages, DIBELS appeared to be the most difficult with a mean score of 86.5 words per minute. EasyCBM was the easiest with a mean score of 95.6 words per minute.

The regression analysis suggested the three sets of ORFs equally predicted the state reading test. Table 5 summarizes these results. Similar to results from our high poverty sample, this sample seemed to suggest DIBELS was the most difficult set of passages, with a mean reading fluency score of 153.2 words per minute. Aimsweb seemed to be the easiest, with a mean fluency rate of 168.0 words per minute.

Summary and discussion

This research seeks to answer practical questions regarding optimal ORF administration. In terms of number of probes used, results of study one suggest that the median of three ORFs may be preferable, although relatively little accuracy appears to be lost by using the mean of two. In study two, findings suggest that 30 s ORF probes may be comparable to 60 s ORF probes. These findings suggest a potential for saving significant administration time by districts moving to three 30 s probes without compromising accuracy of the measure. For example, a medium-sized school district containing 10,000 students in grades K-8 that administers ORFs three times a year would save approximately 250 total testing hours each year. Districts that consider employing three 30 s probes will further minimize variability and reduce test administration time. Using the median of three administrations may also help to control for any undesirable variance associated with passage difficulty. These findings, however, should be interpreted with caution, as future research is needed to further validate the effectiveness of these claims in larger samples and more grade levels.

In terms of relative effectiveness, results of study three suggest that all three ORF assessment systems performed similarly. Each predicted the end-of-year reading tests with similar accuracy. DIBELS, however, was consistently more difficult than Aimsweb and EasyCBM. This may be due to a passage-level effect and may become insignificant over a longer progress-monitoring timeframe. Without more information about the quantifiable differences in passage difficulty, it is prudent for educators to use the norms associated and published with each CBM family.

Table 5
Descriptive statistics and variance explained in end-of-year state test.

	Mean wpm	SD	Adjusted r^2
Aimweb ($n = 67$)	168.0	41.6	0.41*
EasyCBM ($n = 63$)	158.2	34.0	0.41*
DIBELS ($n = 67$)	153.2	35.2	0.40*

* $p < 0.001$.

Limitations

Multiple limitations are associated with the results of the current studies. First, out of convenience, we constrained our data collection to upper elementary grades (5th and 6th grades) located in the Western United States, which limits the generalizability of our findings. Ultimately, a large-scale study will need to examine the relative passage difficulty across all alternate forms across all grades. Second, this research focused on the use of ORF in relation to school-wide screening and thus may not be generalizable to other ORF uses such as progress monitoring. A third limitation, which applies to study one, is that our sample of students screened and found to be in need of assistance did receive assistance in addition to their regular classroom instruction. Thus the scores of this group of students may have changed differentially during the study period when compared to their peers who did not receive this additional instruction. Although this is a common confounding factor in educational research, we recognize a general lack of scientific rigor in this regard.

Future research should examine the degree to which the conclusions from our findings are consistent across grades in larger, more diverse samples. In addition, future research should examine the psychometric impact of redesigning ORF passages specifically for 30 s probes, as this has the potential to save school districts substantial instructional time.

In practice

This series of studies was rooted in practice and has the potential for direct implications in the field. Based on these results, we conclude that administering three 30 s ORF probes and using the median score to represent the *true score* is worthy of serious consideration for practitioners. Additionally, since each probe appears to predict state reading scores with the same accuracy, schools and districts could be equally justified in selecting and using any of the large CBM assessment systems studied here: Aimsweb, EasyCBM, or DIBELS.

References

Alonzo, J., & Tindal, G. A. (2009). *Alternate form and test-retest reliability of EasyCBM reading measures (Technical Report No. 0906)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Ardoin, S. P., & Christ, T. J. (2009). Curriculum-based measurement of oral reading: Estimates of standard error when monitoring progress using alternate passage sets. *School Psychology Review, 38*, 266–283.

Ardoin, S. P., Witt, J. C., Suldo, S. M., Connell, J. E., Koenig, J. L., Resetar, J. L., et al. (2004). Examining the incremental benefits of administering a maze and three versus one curriculum-based measurement reading probes when conducting universal screening. *School Psychology Review, 33*, 218–233.

Barth, A. E., Stuebing, K. K., Fletcher, J. M., Cirino, P. T., Romain, M., Francis, D., et al. (2012). Reliability and validity of oral reading fluency median and mean scores among the middle grade readers when using equated texts. *Reading Psychology, 33*(1–2), 133–161.

California Department of Education. (2002). Test Blueprints Available at http://www.cde.ca.gov/ta/tg/sr/blueprints.asp.

California Department of Education. (2009). *California standards tests technical report spring 2008 administration*. Princeton, NJ: Educational Testing Service.

Children's Educational Services. (1987). *Standard reading passages: Measures for screening and progress monitoring*. Minneapolis, Minnesota: Author.

Christ, T. J. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimates of standard error of slope to construct confidence intervals. *School Psychology Review, 35*, 128–133.

Crawford, L., Gerald, T., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment, 7*(4), 303–323.

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

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

Fuchs, L. S., & Fuchs, D. (2002). Curriculum-based measurement: Describing competence, enhancing outcomes, evaluating treatment effects, and identifying treatment nonresponders. *Peabody Journal of Education, 77*(2), 64–84.

Fuchs, L., Fuchs, D., Hosp, M., & Hamlett, C. (2003). The potential for diagnostic analysis within curriculum based measurement. *Assessment for Effective Intervention, 28*(3–4), 13–22.

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.

Fuchs, L., Tindal, G., & Deno, S. (1981). *Effects of varying time domain and sample duration on technical characteristics of daily measures in reading*. Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.

Good, R. H., & Kaminski, R. A. (1996). Assessment for instructional decisions: Toward a proactive/prevention model of decision-making for early literacy skills. *School Psychology Quarterly, 11*(4), 326–336.

Good, R. H., Kaminski, R. A., Smith, S., & Bratten, J. (2001). *Technical adequacy of second grade DIBELS oral reading fluency passages (Technical Report No. 8)*. Eugene, OR: University of Oregon.

Hasbrouck, J. E., & Tindal, G. (1992). Curriculum-based oral reading fluency norms for students in grades 2 through 5. *Teaching Exceptional Children,* 41–44, (spring).

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

Helwig, R., Heath, B., Tindal, G. (2000). *Predicting middle school mathematics achievement using practical and efficient measurement instruments*. Behavioral Research and Teaching, Eugene, OR: Retrieved from http://brt.uoregon.edu/publications_archive.htm.

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2007). *The ABCs of CBM: A practical guide to curriculum-based measurement*. New York: Guilford Press.

Howe, K. B., & Shinn, M. M. (2002). *Standard reading assessment passages (RAPs) for use in general outcome measurement: A manual describing development and technical features*. Eden Prairie, MN: Edformation.

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

MacMillan, P. D., & Fewster, S. (2002). School-based evidence for the validity of curriculum-based measurement of reading and writing. *Remedial and Special Education, 23*(3), 149–156.

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: Assessment special children* (pp. 18–78). New York, NY: The Gilford Press.

Marston, D., & Deno, S. L. (1987). *Tests of oral reading fluency: Measures for screening and progress monitoring in reading*. Minneapolis, MN: Children's Educational Services Inc.

Oregon Department of Education. (2007). *2006–2007 Technical report; Oregon's statewide assessment system, reliability and validity* (vol. 4, pp.).). Salem, OR: Author.

Reschly, A., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology*, 427–469.

Shinn, M. R., & Bamonto, S. (1998). Advanced application of curriculum-based measurement: "Big Ideas" and avoiding confusion. In M. R. Shinn (Ed.), *Advanced application of curriculum-based measurement* (pp. 1–31). New York, NY: The Gilford Press.

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 2*, 420–428.

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

Tindal, G. A., & Marston, D. B. (1996). Technical adequacy of alternative reading measures as performance assessments. *Exceptionality, 64*, 201–230.

Luke Duesbery, San Diego State University. Dr. Luke Duesbery began his professional career as a 4th grade teacher of gifted and talented students. Since then he has taught elementary and middle school levels, gifted, regular, and special education. He is a licensed teacher and administrator. During his teaching career his focus has been on the integration of modern technologies in teaching. He has presented at local, state, and national conferences. His research interests include educational assessment and measurement in large-scale assessment, curriculum based measurement, the integration of technology in teaching, testing, and learning, and the instructional implications of data graphics assessment.

Jenelle Braun-Monegan, Walden University & Linn-Benton-Lincoln Education Service District. Currently Dr. Braun-Monegan is a school psychologist for the Linn–Benton–Lincoln Education Service District and a contributing faculty member for Walden University. Dr. Braun-Monegan's previous experience includes 3 years as a program specialist in the Beverly Hills Unified School District where she worked under the Director of Special Education to improve the district's special education service delivery. Dr. Braun-Monegan also worked for the evaluation division of a non-profit organization primarily conducting curriculum audits of school districts and evaluations of large statewide grants. Dr. Braun-Monegan received her Ph.D. from the University of Oregon in 2007.

Jacob Werblow, Central Connecticut State University. Dr. Jacob Werblow, is a former 6th grade teacher is one of the largest, urban middle schools in the State of California. Having completed his Ph.D. from the University of Oregon, Jacob is also a licensed teacher and administrator. From 2009 to 2011, Jacob helped the New Britain Public School system receive over $100,000 in external grants. The emphasis of Jacob's research focuses on school equity and effectiveness, student success, and curriculum based measurement. Jacob has published articles in academic journals, such as: The High School Journal, American Secondary Education, Principal, among others.

Drew Braun, Bethel School District. Dr. Drew Braun has been the Director of Instruction for the Bethel School District for 9 years. In that capacity he has incorporated general education, special education, Title, and English Language Development into one department. A priority of Bethel's combined Instruction Department is to coordinate instructional approaches among all programs and to implement research-based practices for all students. Blending the programs has made possible the use of a "systems approach prevention model" that focuses on student learning as a continuum, with multiple and integrated services available to students on an as needed basis.