AMERICAN PSYCHOLOGICAL ASSOCIATION
DIVISION of SCHOOL PSYCHOLOGY Science. Practice. Policy.

School Psychology

© 2020 American Psychological Association
ISSN: 2578-4218

2021, Vol. 36, No. 1, 47–59
http://dx.doi.org/10.1037/spq0000415

Evidence for Automated Scoring and Shorter Passages of CBM-R in Early Elementary School

Joseph F. T. Nese
University of Oregon

Akihito Kamata
Southern Methodist University

Curriculum-based measurement of oral reading fluency (CBM-R) is widely used across the United States as a strong indicator of comprehension and overall reading achievement, but has several limitations including errors in administration and large standard errors of measurement. The purpose of this study is to compare scoring methods and passage lengths of CBM-R in an effort to evaluate potential improvements upon traditional CBM-R limitations. For a sample of 902 students in Grades 2 through 4, who collectively read 13,766 passages, we used mixed-effect models to estimate differences in CBM-R scores and examine the effects of (a) scoring method (comparing a human scoring criterion vs. traditional human or automatic speech recognition [ASR] scoring), and (b) passage length (25, 50, or 85 words, and traditional CBM-R length). We also examined differences in word score (correct/incorrect) agreement rates between human-to-human scoring and human-to-ASR scoring. Our results indicated that ASR can be applied in schools to score CBM-R, and that scores for shorter passages are comparable to traditional passages.

Impact and Implications

The study found evidence that automatic speech recognition (ASR) can be used in schools to score curriculum-based measurement of reading (CBM-R) for consequential assessment and that shorter passages can be used in CBM-R assessment. These findings have implications for the opportunity cost of CBM-R administration. A computerized CBM-R system that incorporates ASR and shorter passages can (a) reduce human administration errors by standardizing administration setting, delivery, and scoring; and (b) reduce the cost of administration by allowing concurrent small-group or whole-classroom testing.

Keywords: oral reading fluency, curriculum-based measurement, automatic speech recognition

Oral reading fluency is the academic construct most often assessed as part of a multitiered system of supports (MTSS) model (Shapiro, Keller, Lutz, Santoro, & Hintze, 2006; Speece, Case, & Molloy, 2003). As such, measures of oral reading fluency are used across the United States, particularly in early elementary grades, to universally screen for students at risk of low reading proficiency, to monitor progress of students receiving reading intervention, and to predict year-end performance on state reading tests. Despite such prevalent use and functional application, the traditional assessment of oral reading fluency has practical limitations tied to construct irrelevant variance, including administration errors (e.g., Cummings, Biancarosa, Schaper, & Reed, 2014; Reed & Sturges, 2013) and large measurement error (e.g., Christ & Silberglitt, 2007; Christ, Zopluoglu, Long, & Monaghen, 2012; Poncy, Skinner, & Axtell, 2005), both of which affect score validity. Automatic speech recognition (ASR) can be used to score oral reading fluency assessments to help minimize or eliminate administration errors. This approach, coupled with the administration of several shorter passages read in their entirety, as opposed to one long passage read for 60 s, is part of a larger solution to help reduce the large measurement error associated with traditional oral reading fluency assessment.

The primary focus of this study is the problem of construct irrelevant variance in oral reading fluency data, some of which comes from human factors (here, administration errors), and some of which comes from passage characteristics (here, passage length). The purposes of this study are twofold. First, we compare scores obtained using both a traditional human score and an ASR score to those obtained by expert curriculum-based measurement

This article was published Online First November 5, 2020.

Joseph F. T. Nese, Behavioral Research and Teaching, University of Oregon; Akihito Kamata, Department of Education Policy and Leadership and Department of Psychology, Southern Methodist University.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140203 to the University of Oregon. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Correspondence concerning this article should be addressed to Joseph F. T. Nese, Behavioral Research and Teaching, University of Oregon, 175 Lokey Education 5262, Eugene, OR 97403-5262. E-mail: jnese@uoregon.edu

of oral reading fluency (CBM-R) scorers using recorded audio files (three scores in total). Second, we evaluate whether new, shorter oral reading fluency passages can be substituted for traditional, longer passages.

Traditional CBM-R

Research devoted to assessing and validating oral reading fluency and studying applied systems of assessments used in schools predominantly relates to CBM-R. CBM in general is designed to measure students' academic status and growth so the effectiveness of instruction may be evaluated (Deno, 1985), and there is strong theoretical support for CBM-R as an essential part of measuring reading proficiency (National Institute of Child Health & Human Development, 2000). Assessing reading fluency is critical because the skill functions as an indicator of comprehension and overall reading achievement (Alonzo, 2016; Fuchs, Fuchs, Hosp, & Jenkins, 2001; Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003; Nese, Park, Alonzo, & Tindal, 2011; Pinnell et al., 1995; Wood, 2006).

In traditional CBM-R administration, students are given 1 min to read as many words as possible in a grade-level text while a trained assessor follows along and indicates on a scoring protocol each word the student reads incorrectly (Wayman, Wallace, Wiley, Tichá, & Espin, 2007). If a student pauses for more than 3 s, the assessor prompts the student to continue and marks the word as read incorrectly. Student self-corrections are not marked as errors, but word omissions are. After 1 min, the assessor calculates the fluency score as words correct per minute (WCPM) by subtracting the number of incorrectly read words from the total number of words read.

Errors in Traditional CBM-R Administration

The opportunity for error in the CBM-R administration process is exceedingly high, and has been well-documented in the literature (Christ, 2006; Christ & Silberglitt, 2007; Colón & Kranzler, 2006; Derr-Minneci & Shapiro, 1992; Munir-McHill, Bousselot, Cummings, & Smith, 2012; Reed, Cummings, Schaper, & Biancarosa, 2014; Reed & Sturges, 2013). In an examination of assessment fidelity in the administration, scoring, and interpretation of CBM-R tests, Reed and Sturges (2013) reported that 8% of 589 tests were flagged for significant abnormalities that could not be corrected, including forgetting to start the timer, not stopping the student or circling the last word when the timer sounded, and other errors. In addition, approximately 91% of the remaining 532 tests had correctable errors, including counting insertions as errors, miscounting the number of errors, and miscalculating the WCPM.

A literature synthesis of 46 studies on assessment fidelity in reading intervention research concluded that despite the variety of systematic errors that are possible when testing students, assessment integrity data were rarely reported in the examined literature (Reed et al., 2014). But research has shown that CBM-R scores are sensitive to characteristics of the administrator and location of the administration (Derr-Minneci & Shapiro, 1992), as well as the delivery of directions (Colón & Kranzler, 2006). Distractions and inconsistent directions result in deteriorated or invalid CBM-R data (Christ, 2006; Christ & Silberglitt, 2007), and administration settings and procedures are as important to valid CBM-R scores as passage equivalence (Christ, 2006). Despite these findings, a study of 160 schools showed that only 27% of schools reported that all examiners had received formal instruction in the standardized assessment procedures, and even worse, 24% of schools reported that no formalized training opportunities were provided to their assessors (Munir-McHill et al., 2012). Another recent study examined the variance in 11,777 CBM-R scores of 2,960 students and found 16% of the variance was attributed to assessors (Cummings et al., 2014).

The Large Standard Error (*SE*) of Traditional CBM-R Scores

Traditional CBM-R scores are generally associated with large measurement error, or large *SE* (Ardoin & Christ, 2009; Christ & Silberglitt, 2007; Christ et al., 2012; Poncy et al., 2005), which can affect the interpretations and consequences of the assessment results as teachers use CBM-R fluency scores to screen for students at risk of poor reading outcomes and monitor student progress to inform instruction (Christ & Coolong-Chaffin, 2007). Along with a psychometric modeling approach (Kara, Kamata, Potgieter, & Nese, 2020; Potgieter, Kamata, & Kara, 2017), having students read several shorter passages may be part of a larger solution to reduce the large *SE* associated with traditional CBM-R fluency scores and increase score reliability (Nese & Kamata, 2020). Thus, in this study we explore whether shorter passages (25 to 85 words) read in their entirety function comparably to traditional CBM-R passages (about 250 words) read for 60 s.

Application of ASR for CBM-R Assessment

A computerized CBM-R procedure has the potential to minimize or eliminate administration errors by standardizing the delivery, setting, and scoring (e.g., timing the reading for exactly 60 s, correctly marking the last word read in the allotted time, correctly calculating the number of words read correctly, and recording the correct WCPM score in the database). In addition, small groups or an entire classroom can be assessed simultaneously in only a few minutes so that a single educator can monitor the integrity of the environment for a group of students. Such an approach would have the benefit of both reducing the cost and time associated with collecting CBM-R data and limiting construct irrelevant variance attributable to differences in CBM-R administration and scoring practices.

There is a small body of research exploring the use of ASR for CBM-R assessment, all of which has shown a strong relation between ASR and expert human WCPM scores, with correlations ranging from .86 to .99 (Bernstein, Cheng, Balogh, & Rosenfeld, 2017; Bolaños, Cole, Ward, Borts, & Svirsky, 2011; Zechner, Evanini, & Laitusis, 2012; Zechner, Sabatini, & Chen, 2009). But there are several limitations to this research, including the fact that although the number of studies in this area is growing, it remains rather small. In addition, the correlations reported above represent the relation between ASR and expert human assessors' WCPM scores of audio recordings, and not between ASR and traditional human assessors' CBM-R scores conducted in real time and in school settings. That is, in the ASR studies conducted thus far, trained assessors listen to recordings of readings on headphones with high-quality acoustic signal and replay as needed to refine the scoring, and the scores are thus not subject to some of practical limitations of traditional CBM-R administration previously described. Although the results help provide important validity evidence for ASR scoring of WCPM, no prior studies compare ASR, traditional real-time CBM-R scores, and scores of expert scorers obtained using recordings of the students reading aloud. Finally, no prior study has reported the rates of word score agreement reached by people scoring from recorded audio as compared to ASR or traditional, in-person human CBM-R scoring.

Current Study

Data for the present study were drawn from the larger Computerized Oral Reading Evaluation (CORE; https://jnese.github.io/core-blog/) project which focused on the development and validation of a computerized CBM-R assessment system to address the practical and methodological limitations of oral reading fluency assessment. As part of the CORE project, passages of different lengths (i.e., 25, 50, and 85 words) were developed and administered to students in Grades 2, 3, and 4, along with traditional CBM-R passages. Then, passages of all four lengths were scored in three different ways. First, trained members of the research team scored each passage as it was being read, as is done in traditional CBM-R assessments. Second, the computer system used ASR to score each passage. Finally, researchers with expertise in CBM-R administration and scoring listened to audio recordings of the CBM-R assessments in a quiet environment with the ability to pause and rewind to verify their scoring decisions. These scores served as the criterion scores in this study. Our research questions were as follows.

(1) Are there statistically significant differences in the time recorded for a student to read a passage, as measured by the traditional human scorer compared to the ASR?

(2) Are there statistically significant differences in students' oral reading fluency scores, measured in WCPM, as a function of the three scoring methods, passage length, or the interaction of those two factors?

(3) Are there meaningful differences in word score agreement rates as a function of the three scoring methods?

Method

This study was conducted in the springs of 2015 (Year 1) and 2016 (Year 2) in Oregon, with institutional review board approval. All CORE passages were developed in 2014: half were then designated for use in each of the two subsequent years. Year 2 was a replication of Year 1, differing only in the sample of participants and the CORE passages used. That is, Year 2 used the second half of the developed CORE passages but the same traditional CBM-R passages.

Sample

The original data included 937 students. All students in Grades 2 through 4 at participating schools were invited to participate such that the sample would be representative, to the extent possible, of typically developing students across reading proficiency levels. Table 1 presents the demographic data available from each school, as well as the accompanying National Center for Education Statistics (NCES) 2015–2016 school Common Core of Data (http://nces.ed.gov/ccd/pubagency.asp). We did not require systematic student demographic information from each school, and thus missing data across student demographic variables (8% to 23%) make generalizations difficult. According to 2015–2016 NCES school data, the populations of the four schools ranged from 423 to 523 students, approximately half of whom were students in Grades 2 through 4. Two school locales were classified as Town: Distant, and two as Suburb: Midsize (for more information, see https://nces.ed.gov/ccd/commonfiles/glossary.asp). All four schools received Title I funding, and the percentage of students receiving free or reduced lunch ranged from 67% to 85%. The ethnic/race majority for all schools was White (63% to 89%).

These students provided a total of 14,711 audio files/passages read. We removed 934 audio files (6.3%) that were either shorter than 5 s (with the assumption that it was not reasonable any passage could be read in that time) or longer than 90 s (with the assumption that students were generally not allowed to read that long; see the Administration section below). We removed an additional 11 audio files (0.1%) that produced unusable data due to sound quality. Thus, the final analytic sample consisted of 902 students (262 Grade 2, 325 Grade 3, and 315 Grade 4) from four schools, and 13,766 passages read.

Measures

CORE CBM-R passages. Each CORE passage is an original work of narrative fiction that follows the story grammar of English language short stories, with a main character and a clear beginning, middle, and end (http://bit.ly/core_2E8iZDF). To reduce construct-irrelevant variance associated with different authors' voice and style, the author of the CORE passages was part of the team that authored the easyCBM traditional CBM-R passages used in this study. Apart from the passage length requirements, the CORE passages were written to similar specifications as the easyCBM passages. Three different lengths of CORE passages were written: short = 25 words, medium = 50 words, and long = 85 words (all ± 5 words). Ultimately, 330 passages were written: 110 at each of Grades 2–4, with 20 long passages, 30 medium passages, and 60 short passages for each grade. Administration instructions were to allow students to read the CORE passages in their entirety, and to record the elapsed time in seconds.

Traditional CBM-R passages. We administered the easyCBM (Alonzo, Tindal, Ulmer, & Glasgow, 2006) oral reading fluency measures as the traditional CBM-R assessments for the purpose of comparison to CORE passages for construct validity. The easyCBM CBM-R measures have demonstrated features of technical adequacy that suggest they are sufficient to meet the needs as the comparative example of an existing traditional CBM-R system (Anderson et al., 2014). Following standard administration protocols, students were given 60 s to read the traditional CBM-R passages.

Procedure

Human scorer training. We recruited and trained human scorers to administer and score all study passages. Year 1 traditional human scorers included 14 individuals: five with doctoral degrees in education, five graduate students in a college of edu-

Table 1
School Demographic Percentages Based on Study Data (Top) and NCES 2015–2016 Common Core of Data (Bottom)

Demographics	School A	School B	School C	School D
Approximate % of sample	37%	7%	35%	25%
Sex				
Female	35%	53%	44%	41%
Male	42%	35%	33%	51%
Missing	23%	12%	23%	8%
Students with disability				
No	64%	75%	71%	78%
Yes	13%	13%	6%	14%
Missing	23%	12%	23%	8%
Ethnicity				
Hispanic/Latino	12%	3%	16%	23%
Not Hispanic/Latino	65%	85%	61%	69%
Missing	23%	12%	23%	8%
Race				
American/Alaskan Native	1%	3%	2%	3%
Asian	1%	0%	3%	1%
Black	0%	4%	1%	2%
Multirace	0%	0%	4%	10%
Pacific Islander	1%	0%	0%	1%
White	73%	81%	67%	74%
Missing	23%	12%	23%	8%
EL				
No	70%	88%	70%	78%
Yes	6%	0%	7%	14%
Missing	23%	12%	23%	8%
NCES 2015–16 data				
Total Students	473	423	523	442
% Grades 2–4	55%	55%	52%	50%
% Female	46%	47%	51%	45%
% White	77%	89%	63%	63%
% FRL	80%	73%	67%	85%
Title I	Yes	Yes	Yes	Yes
Locale description	Town: Distant	Town: Distant	Suburb: Midsize	Suburb: Midsize

Note. NCES = National Center for Education Statistics; EL = English learner.

cation, one with a master's degree in the field of education, and three university staff. Year 2 traditional human scorers included 21 individuals: seven undergraduate students, seven graduate students in a college of education, four with doctoral degrees in education, two with a master's degree in the field of education, and two university staff (nine also served as traditional human scorers in Year 1). A subset of the traditional human scorers also served as human criterion scorers. During the training sessions, the scorers practiced scoring passages that were read aloud by a training leader, with planned errors, omissions, self-corrections, and disfluent reading, and the reliability of the human scoring was calculated using Cohen's (1960) kappa (κ) procedure, where κ was calculated as the word agreement between each individual and the true scores, adjusted for agreement by chance. That is, for each scorer, each word in the passage was scored as read correctly/incorrectly, and κ was the pairwise comparison to the planned errors for the passage. The 14 Year 1 κ coefficients were all above .90, and the 21 Year 2 κ coefficients were all above .87.

Administration. Students were assessed online, via laptops in a one-to-one administration setting. Students were seated in front of a computer and given standardized introduction and task instructions. Students wore headphones with an attached noise-cancelling microphone. While students read, an assessor followed on a paper copy, marking words the student omitted or read incorrectly, and also marking the last word read. If a student failed to produce any words in the first 10 s and it was clear they were a nonreader, the assessment was terminated and the student was excluded from the study. In Year 2 only, if a student read below the 10th percentile on the first traditional CBM-R passage (Saven, Tindal, Irvin, Farley, & Alonzo, 2014), then for all remaining CORE passages, they were instructed to read only for 30 s, and the last word read after 30 s was marked by the assessor. We instituted this rule so as not to stress exceptionally poor readers by having them read for an excessive amount of time, while also creating an avenue by which they could participate in the study.

In all cases, the first passage presented was the grade-level traditional CBM-R passage (easyCBM), which was read for 60 s, after which the student read each assigned CORE passage in its entirety. In Year 1, students read 18 CORE passages (3 long, 5 medium, and 10 short), and in Year 2, students read 14 CORE passages (2 or 3 long, 3 or 4 medium, and 7 or 8 short). (Note that the initial intent of the larger project was that these groupings by length were each approximately 250 words in total, equivalent in length to traditional CBM-R passages.) CORE passages were randomly assigned at the student level prior to assessment administration and were grouped to control for potential fatigue and order effects. Students read an average of 17.2 (*SD* = 2.3) passages in Year 1 (average total reading time about 7.5 mins) and 13.3 (*SD* = 2.2) in Year 2 (average total reading time about 6.1 mins).

Scoring and time recorded. Each passage read by each participant received three WCPM scores (described below). The WCPM scores were the number of correctly read words divided by the recorded passage reading time (in seconds) and multiplied by 60, so that all fluency scores were on the same scale regardless of the time it took to read the passage. Table 2 shows the descriptive passage data.

Human criterion score. The recorded audio files were scored by the trained assessors at a later date and in a different setting from which the data were collected. Assessors wore headsets and listened to each recorded audio file (with the ability to rewind, replay, and adjust audio), using the same scoring rules as the traditional procedures. These scores are considered the criterion to which the traditional CBM-R and ASR scores are compared. The total time for a student to read a passage was not recorded by the human criterion scorer. Instead, we used the time recorded by the ASR (described below) in calculating the human criterion WCPM scores.

Traditional human score. As in traditional CBM-R administration, students read while the trained assessor followed along and marked each word read incorrectly (Wayman et al., 2007). If a student paused for more than 3 s, the assessor marked the word as read incorrectly and prompted the student to continue reading without providing the correct word, which is the only deviation from traditional administration. The traditional human scorer recorded the time it took a student to read a passage (first word read to last word read). For CORE passages, this was the time it took a student to read each passage in its entirety. For traditional CBM-R passages, this time was 60 s. The traditional human scorer recorded and entered the number of words read correctly, the last word read, and the time it took to read from the first word read to the last word read; however, the traditional human scorers were not asked to calculate the WCPM score by hand. That is, the traditional human WCPM scores were calculated automatically by computer: dividing the number of words read correctly by the recorded time and multiplying the result by 60.

Automatic speech recognition (ASR) score. The ASR engine scored each audio recording file, scoring each word as read correctly or incorrectly (just like both human scorers), and recording the time (in centiseconds) it took a student to read each word, and the time (in centiseconds) between words. For the ASR time score, we calculated the time it took a student to read a passage (first word read to last word read) by summing the time it took to read each word as well as the timed silence between each word. For CORE passages, the time was the total time it took a student to read the passage in its entirety. For traditional CBM-R passages, the time was stopped at the last complete word read 60 s after the first word read, and thus the recorded time for CBM-R passages was generally 60 s or slightly less.

Bavieca, an open-source speech recognition toolkit, was the ASR applied in this study (http://www.bavieca.org/). Bavieca uses continuous density hidden Markov models and supports maximum likelihood linear regression, vocal tract length normalization, and discriminative training (maximum mutual information). It uses the general approach of many state-of-the art speech recognition systems: a Viterbi Beam Search used to find the optimal mapping of the speech input onto a sequence of words. The score for a word sequence was calculated by interpolating language model scores and acoustic model scores. The language model assigned probabilities to sequences of words using trigrams (where the probability of the next word is conditioned on the two previous words) and was trained using the CMU-Cambridge LM Toolkit (Clarkson & Rosenfeld, 1997). Acoustic models were clustered triphones based on Hidden Markov models using Gaussian mixtures to estimate the probabilities of the acoustic observation vectors. The system used filler models to match the types of disfluencies found in applications.

Analyses

The unit of observation for the data was either reading time recorded or WCPM for a specific student for a specific passage. Because each student read multiple passages, there were repeated

Table 2
Mean (SD) WCPM for Scoring Methods, and Counts of Passages, Students, and Audio Recordings by Grade and Passage Length

	Recording		ASR		Traditional		Count		
Passages	*M*	*SD*	*M*	*SD*	*M*	*SD*	Unique passages	Students	Passages read
Grade 2									
Traditional CBM-R	100.7	(32.3)	83.7	(27.8)	90.7	(31.8)	1	224	224
Long	95.7	(33.7)	91.6	(33.7)	95.4	(39.4)	18	242	537
Medium	91.5	(38.5)	86.7	(37.1)	91.3	(43.3)	29	255	988
Short	84.6	(42.3)	81.5	(41.3)	89.5	(47.6)	59	262	2,104
Grade 3									
Traditional CBM-R	119.2	(31.1)	101.9	(28.5)	115.5	(35.5)	1	302	302
Long	111.5	(34.9)	106.4	(35.1)	115.0	(43.0)	19	316	816
Medium	108.5	(36.4)	104.8	(36.0)	113.5	(39.7)	30	319	1,350
Short	112.2	(42.6)	108.9	(42.1)	119.3	(46.1)	56	323	2,540
Grade 4									
Traditional CBM-R	124.2	(32.6)	106.0	(30.4)	123.1	(37.9)	1	300	302
Long	128.5	(39.4)	124.4	(38.6)	132.1	(42.1)	20	309	860
Medium	132.4	(42.2)	128.7	(42.5)	137.9	(46.0)	27	311	1,229
Short	132.1	(44.7)	128.5	(43.9)	141.4	(48.8)	58	312	2,514

Note. WCPM= words correct per minute; ASR = automatic speech recognition; CBM-R = curriculum-based measurement of oral reading fluency.

measures for each student. However, students were not nested within passages, or vice versa, because a set of observations for each student was not from the same set of passages as all other students. Thus, the data had a cross-classified structure, where students and passages were cross-classified. Therefore, to address Research Questions (RQ) 1 and 2, we applied a linear mixed-effect model (LMM) by treating both students and passages as random effects, separately for each of Grades 2 through 4 and each outcome, time recorded and WCPM (3 grades × 2 outcomes resulted in 6 models).

Based on our final models, we calculated pairwise differences (in time and WCPM) based on the estimated marginal means of the mixed-effect models. We set threshold p value $< .01$ to determine statistical significance and applied the Bonferroni adjustment to control the family-wise error rate for pairwise comparisons. To assist the interpretation of the pairwise comparisons, we also report Cohen's *d* (Cohen, 1988) as the effect size (*ES*) of the estimated differences. To address Research Question 3, we calculated the percent of words in each passage that were scored in agreement (as correct or incorrect) between the recording criterion and the ASR or traditional scores. Lastly, Williamson, Xi, and Breyer (2012) provided a framework and guidelines to evaluate automated scoring for consequential assessment, and we use some of their performance criterion to contextualize the results.

Analyses were conducted and figures were created in the R programming environment (R Core Team, 2019) with the following packages: broom (Robinson & Hayes, 2018), devtools (Wickham, Hester, & Chang, 2018), emmeans (Lenth, 2019), esvis (Anderson, n.d.), here (Müller, 2017), lme4 (Bates, Mächler, Bolker, & Walker, 2015), patchwork (Pedersen, 2019), and tidyverse (Wickham, 2017).

Results

We began by comparing unconditional models with and without random effects for student and passage. For each outcome and grade, deviance test results showed that the model with random effects for both student and passage statistically improved the model fit compared to models with a random effect for either student or passage (results available upon request). Most of the variance (54% to 72%) for the baseline time models was between passages (Table 3), which makes sense given that passage length and time to read a passage are highly and directly related (i.e., it takes more time to read a longer passage). Most of the variance (66% to 74%) for the baseline WCPM models was between students (Table 4).

To the model with random effects for student and passage, we then added fixed-effects for passage length (four levels: traditional CBM-R, short, medium, and long) and scoring method (three levels: human criterion, traditional human, and ASR). For each outcome and grade, deviance test results showed that the addition of an interaction between scoring method and passage length statistically improved the model fit compared to models without that interaction (results available upon request). Thus, our final model for both outcomes and all grades included random effects for student and passage and fixed effects for passage length, scoring method, and their interaction.

RQ 1: Time Recorded to Read Passages

Table 3 shows the results of the final time models, with random effects for student and passage and fixed effects for passage length, scoring method, and the interaction of those two factors. The intercept represents the average time in seconds to read traditional CBM-R passages as recorded by the ASR (Grades 2 = 57.1 s, Grade 3 = 56.9 s, and Grade 4 = 57.3 s). All pairwise comparisons (across grades and passage length) between the ASR and the traditional human times were statistically significant at the adjusted α level of .0025 (.01/4 for four pairwise comparisons; Table 5). The time recorded by the traditional human scorers was greater than the ASR for the traditional CBM-R passages across grades, and less than the ASR for the CORE passages. An examination of the magnitude of the *ES*s showed quite large time differences for

Table 3
Time Outcome: Fixed and Random Effects From Final Mixed-Effects Models of Recorded Time, Grades 2 through 4

	Grade 2			Grade 3			Grade 4		
Fixed effects	Estimate	*SE*	*t*-value	Estimate	*SE*	*t*-value	Estimate	*SE*	*t*-value
Intercept	57.1	3.2	18.0	56.9	3.2	18.0	57.3	2.2	26.6
Long	−2.7	3.2	−0.8	−8.4	3.2	−2.6	−15.3	2.2	−7.0
Medium	−21.5	3.2	−6.7	−27.4	3.2	−8.6	−32.9	2.2	−15.1
Short	−35.6	3.2	−11.2	−41.5	3.2	−13.1	−44.7	2.2	−20.8
Traditional human score	3.9	0.7	5.4	3.6	0.5	7.4	2.8	0.4	7.0
Long: Traditional	−6.4	0.9	−7.4	−5.3	0.6	−9.3	−4.2	0.5	−9.0
Medium: Traditional	−5.8	0.8	−7.1	−4.9	0.5	−9.2	−3.7	0.4	−8.4
Short: Traditional	−5.5	0.8	−7.1	−4.5	0.5	−8.7	−3.5	0.4	−8.4

Random effects	Baseline model	Final model	Baseline model	Final model	Baseline model	Final model
Student	75.0	74.9	50.7	51.0	30.3	30.3
Passage	161.7	9.6	170.9	9.7	140.8	4.5
Residual	60.3	59.3	36.1	35.6	24.5	24.2

Note. Intercept represents the average time in seconds to read traditional CBM-R (curriculum-based measurement of oral reading fluency) passages as recorded by the ASR (automatic speech recognition). The Baseline model represents the variance estimates from the model with no fixed effects, and the Final model represents variance estimates from the model with the fixed effects presented in the upper part of the table.

Table 4
WCPM Outcome: Fixed and Random Effects From Final Mixed-Effects Models of WCPM, Grades 2 Through 4

	Grade 2			Grade 3			Grade 4		
Fixed effects	Estimate	*SE*	*t*-value	Estimate	*SE*	*t*-value	Estimate	*SE*	*t*-value
Intercept	95.6	8.8	10.8	117.9	11.7	10.1	123.9	10.9	11.4
Long	−9.1	8.8	−1.0	−9.5	11.8	−0.8	3.0	11.0	0.3
Medium	−7.6	8.7	−0.9	−10.6	11.7	−0.9	7.6	10.9	0.7
Short	−12.0	8.6	−1.4	−5.8	11.6	−0.5	8.7	10.8	0.8
ASR	−17.0	1.8	−9.2	−17.3	1.7	−10.2	−18.2	1.7	−10.9
Traditional human score	−10.0	1.8	−5.4	−3.7	1.7	−2.2	−1.0	1.7	−0.6
Long: ASR	12.9	2.2	5.9	12.2	2.0	6.1	14.1	1.9	7.2
Medium: ASR	12.2	2.0	6.0	13.6	1.9	7.3	14.4	1.9	7.7
Short: ASR	14.0	1.9	7.2	14.0	1.8	7.8	14.5	1.8	8.2
Long: Traditional human score	9.8	2.2	4.4	7.2	2.0	3.6	4.7	1.9	2.4
Medium: Traditional human score	9.9	2.0	4.8	8.7	1.9	4.6	6.5	1.9	3.5
Short: Traditional human score	14.9	1.9	7.7	10.7	1.8	6.0	10.3	1.8	5.8

Random effects	Baseline model	Final model	Baseline model	Final model	Baseline model	Final model
Student	1,346.4	1,346.5	1,156.3	1,157.2	1,419.5	1,420.1
Passage	71.6	70.8	138.7	130.9	123.6	113.2
Residual	393	382.2	453.1	434.3	448.7	423.3

Note. ASR = automatic speech recognition. Intercept represents the average WCPM (words correct per minute) to read traditional CBM-R (curriculum-based measurement of oral reading fluency) passages as recorded by the human criterion scores. The Baseline model represents the variance estimates from the model with no fixed effects, and the Final model represents variance estimates from the model with the fixed effects presented in the upper part of the table.

the traditional CBM-R passages across grades (about 1.0 *SD*), and medium *ES*s for the CORE passages (about 0.10 to 0.20 *SD*).

RQ 2: WCPM—Scoring and Passage Length Effects

Table 4 shows the results of the final WCPM mixed effects models. The intercept represents the average WCPM to read traditional CBM-R passages as recorded by the human criterion scorers (Grades 2 = 95.6 WCPM, Grade 3 = 117.9 WCPM, and Grade 4 = 123.9 WCPM). Table 6 shows the pairwise differences in estimated WCPM between the criterion human scores and the traditional human or ASR scores, by passage length. All pairwise comparisons (across grades and passage length) between the human criterion and the ASR or traditional human WCPM scores were statistically significant at the adjusted α level of .00125 (.01/8), with the following four exceptions for human criterion–traditional human scores: Grade 2 long and medium passages, and Grades 3 and 4 traditional CBM-R passages (Table 6). Thus, the human-to-human scores were not statistically different for four of the 24 comparisons across grades.

In general, the *ES*s for traditional CBM-R passages ranged from zero to large (0.03 to 0.58), whereas those for the CORE passages ranged from zero to medium (−0.20 to 0.15), meaning that the differences from the human criterion scores were more pronounced, for both the traditional human and the ASR scores, for the traditional CBM-R passages. For the traditional CBM-R passages, both the ASR and the traditional human scores always underestimated the WCPM compared to the human criterion, and the ASR always more so. For all of the CORE passage lengths (long, medium, and short) across grades, the ASR always underestimated, and the traditional human generally overestimated, the WCPM compared to the human criterion scores.

Williamson and colleagues (2012) suggest a limit of 0.15 as an acceptable threshold for the standardized mean score difference between human scores and automated scores. Of the 24 pairwise *ES*s presented in Table 6, only three between the human criterion and the ASR are greater than 0.15 (those for the traditional CBM-R passages in each grade). In addition, one human criterion–traditional human *ES* was greater than 0.15 (Grade 2 traditional CBM-R passage).

Table 5
Pairwise Differences in Time Duration Between ASR and Traditional Scoring Methods

	Grade 2				Grade 3				Grade 4			
ASR–Traditional human score	Estimate	*SE*	*p*-value	*d*	Estimate	*SE*	*p*-value	*d*	Estimate	*SE*	*p*-value	*d*
Traditional CBM-R	−3.9	0.7	<.0025	−1.08	−3.6	0.5	<.0025	−1.04	−2.8	0.4	<.0025	−1.11
Long	2.4	0.5	<.0025	0.15	1.7	0.3	<.0025	0.12	1.4	0.2	<.0025	0.12
Medium	1.8	0.3	<.0025	0.14	1.4	0.2	<.0025	0.14	0.9	0.2	<.0025	0.11
Short	1.5	0.2	<.0025	0.15	0.9	0.2	<.0025	0.13	0.7	0.1	<.0025	0.14

Note. *d* = Cohen's *d* (Cohen, 1988). The Bonferroni adjustment was applied to control the family-wise error rate for pairwise comparisons so the adjusted α level was .01/4 = .0025. ASR = automatic speech recognition; CBM-R = curriculum-based measurement of oral reading fluency.

Table 6
Pairwise Differences in Estimated WCPM Between Scoring Methods by Traditional (easyCBM) and CORE (Long, Medium, Short) Passage Lengths

	Grade 2				Grade 3				Grade 4			
Scoring method comparisons	Estimate	*SE*	*p*-value	*d*	Estimate	*SE*	*p*-value	*d*	Estimate	*SE*	*p*-value	*d*
Traditional CBM-R												
Human criterion–ASR	17.0	1.8	<.00125	0.57	17.3	1.7	<.00125	0.58	18.2	1.7	<.00125	0.58
Human criterion–Traditional human	10.0	1.8	<.00125	0.31	3.7	1.7	.031	0.11	1.0	1.7	.538	0.03
Long												
Human criterion–ASR	4.1	1.2	<.00125	0.12	5.1	1.0	<.00125	0.15	4.1	1.0	<.00125	0.10
Human criterion–Traditional human	0.3	1.2	.817	0.01	−3.5	1.0	<.00125	−0.09	−3.7	1.0	<.00125	−0.09
Medium												
Human criterion–ASR	4.8	0.9	<.00125	0.13	3.7	0.8	<.00125	0.10	3.8	0.8	<.00125	0.09
Human criterion–Traditional human	0.1	0.9	.884	0.00	−5.0	0.8	<.00125	−0.13	−5.4	0.8	<.00125	−0.12
Short												
Human criterion–ASR	3.1	0.6	<.00125	0.07	3.3	0.6	<.00125	0.08	3.7	0.6	<.00125	0.08
Human criterion–Traditional human	−4.9	0.6	<.00125	−0.11	−7.1	0.6	<.00125	−0.16	−9.2	0.6	<.00125	−0.20

Note. CORE = Computerized Oral Reading Evaluation; *d* = Cohen's *d* (Cohen, 1988). The Bonferroni adjustment was applied to control the family-wise error rate for pairwise comparisons so the adjusted α level was .01/8 = .00125. CBM-R = curriculum-based measurement of oral reading fluency; ASR = automatic speech recognition.

Table 7 shows the pairwise differences in estimated WCPM between passage length, by scoring method. None of the WCPM pairwise comparisons (across grades and scoring methods) between the traditional CBM-R passages and the shorter CORE passages were statistically significant at the adjusted α level of .0011 (.01/9). (Note that we do not apply the guidelines provided by Williamson et al., 2012, because we are not comparing human and machine scores here.)

RQ 3: Word Agreement

Table 8 shows the percent of words read in a passage scored in agreement (as correct or incorrect) between the human criterion and the traditional human or ASR scores, as well as the degradation between human-to-human and human-to-ASR agreement. In general, the average agreement rates between the human criterion and the traditional human scores were exceptionally high, ranging from .97 to .99, and the average agreement rates between human criterion and the ASR scores were also high, ranging from .81 to .94 (all but one were above .87). The *SD*s for the human criterion and the traditional human agreement rates (.02 to .05) were 2 to 8 times smaller than those for ASR and human criterion (.08 to .19), indicating much less variance in the former. Figure 1 shows the box plot distributions of passage-level agreement rates by grade and passage length, separately for the human criterion and ASR, and the human criterion and traditional human. The interquartile ranges for the human criterion and traditional human agreement rates are much smaller and generally set near perfect agreement, whereas those interquartile ranges are larger for human criterion and ASR but also set near .88 or above.

Williamson et al. (2012) suggest that degradation between human-to-human and human-to-ASR agreement should not be

Table 7
Pairwise Differences in Estimated WCPM Between the Criterion Human Scores and the Traditional Human or ASR Scores, by Passage Length

	Grade 2				Grade 3				Grade 4			
Passage length comparisons	Estimate	*SE*	*p*-value	*d*	Estimate	*SE*	*p*-value	*d*	Estimate	*SE*	*p*-value	*d*
Recording												
Traditional CBM-R–long	9.1	8.8	0.299	0.15	9.5	11.8	0.423	0.23	−3.0	11.0	0.784	−0.11
Traditional CBM-R–medium	7.6	8.7	0.381	0.25	10.6	11.7	0.364	0.30	−7.6	10.9	0.486	−0.20
Traditional CBM-R–short	12.0	8.6	0.164	0.39	5.8	11.6	0.618	0.17	−8.7	10.8	0.421	−0.18
ASR												
Traditional CBM-R–long	−3.8	8.8	0.665	−0.25	−2.7	11.8	0.819	−0.13	−17.1	11.0	0.119	−0.50
Traditional CBM-R–medium	−4.6	8.7	0.594	−0.08	−3.0	11.7	0.798	−0.09	−22.0	10.9	0.044	−0.56
Traditional CBM-R–short	−2.0	8.6	0.816	0.05	−8.2	11.6	0.481	−0.17	−23.2	10.8	0.032	−0.53
Traditional												
Traditional CBM-R–long	−0.6	8.8	0.943	−0.13	2.3	11.8	0.846	0.01	−7.7	11.0	0.483	−0.22
Traditional CBM-R–medium	−2.3	8.7	0.791	−0.01	2.0	11.7	0.866	0.05	−14.1	10.9	0.198	−0.33
Traditional CBM-R–short	−3.0	8.6	0.729	0.03	−4.9	11.6	0.671	−0.08	−19.0	10.8	0.079	−0.38

Note. *d* = Cohen's *d* (Cohen, 1988). The Bonferroni adjustment was applied to control the family-wise error rate for pairwise comparisons so the adjusted α level was .01/9 = .0011. WCPM = words correct per minute; CBM-R = curriculum-based measurement of oral reading fluency; ASR = automatic speech recognition.

Table 8
Mean (SD) Agreement Rates Between the Human Criterion Word Scores and the Traditional Human or ASR Scores, Aggregated by Grade and Passage Length

	Human criterion and traditional human		Human criterion and ASR		
Passages	*M*	*SD*	*M*	*SD*	Degradation
Grade 2					
Traditional CBM-R	.97	(.03)	.81	(.19)	.16
Long	.98	(.04)	.88	(.16)	.10
Medium	.98	(.05)	.88	(.16)	.10
Short	.97	(.05)	.88	(.14)	.09
Grade 3					
Traditional CBM-R	.98	(.03)	.90	(.13)	.08
Long	.98	(.03)	.92	(.11)	.06
Medium	.98	(.03)	.93	(.09)	.05
Short	.98	(.04)	.93	(.10)	.05
Grade 4					
Traditional CBM-R	.98	(.03)	.92	(.12)	.06
Long	.98	(.02)	.93	(.09)	.05
Medium	.99	(.02)	.94	(.08)	.05
Short	.98	(.03)	.94	(.09)	.04

Note. ASR = automatic speech recognition; CBM-R = curriculum-based measurement of oral reading fluency.

more than .10. Of the 12 agreement rate comparisons in Table 8 (4 passage lengths × 3 grades), only one was larger than .10, providing evidence for adequate degradation.

Discussion

Although CBM-R is a good indicator of reading proficiency, the assessment has several inadequacies that could be improved upon, including errors in administration (Colón & Kranzler, 2006; Cummings et al., 2014; Derr-Minneci & Shapiro, 1992; Munir-McHill et al., 2012; Reed et al., 2014), high cost in time and resources of administration and implementation (Hoffman, Jenkins, & Dunlap, 2009), and large standard error of measurement (Ardoin & Christ, 2009; Christ & Silberglitt, 2007; Poncy et al., 2005). The present study compared human and automated scoring methods and passage lengths to examine whether ASR can adequately score CBM-R assessments and whether shorter passages can adequately replace traditional CBM-R passages.

Time Recorded to Read Passages

In response to our first research question, we found statistically significant differences between traditional human scorers and ASR in the time recorded for students to read passages.

All pairwise comparisons between traditional human and ASR scoring were statistically significant (Table 5). These differences favor the ASR, under the assumption that the ASR timings, which recorded the time in centiseconds to read each word as well as the silences between words, were very near precise. On average, the time recorded by traditional human scorers was about 1.0 *SD* greater than the ASR for the traditional CBM-R passages across grades (about 3 to 4 s), and a little more than 0.10 *SD* for the shorter CORE passages (about 1 to 2 s). These differences can affect a CBM-R score by as much as seven WCPM. This finding suggests that humans were not as accurate as the ASR at recording the time to read a passage, particularly the 60 s to read a traditional CBM-R passage, a finding that has been previously documented in the literature (Reed & Sturges, 2013). The ASR times were not infallible across the 13,766 passages read, but in general and under reasonable assumptions, the ASR timing appears to be more accurate than traditional human CBM-R times, which are susceptible to many different types of human errors (Christ, 2006; Reed & Sturges, 2013; Reed et al., 2014). These findings support the application of ASR in schools to score CBM-R assessments, as timing students' readings directly affects WCPM fluency scores. A more accurate time should result in a more accurate WCPM score, which is important when those scores are used for consequential decisions about reading intervention and response to instruction.

WCPM—Scoring Methods

In response to our second research question, we found statistically significant differences in WCPM scores between the human criterion and both ASR and traditional CBM-R scoring (see Table 6). On average, both the ASR and the traditional human scores underestimated WCPM on traditional CBM-R passages compared to the human criterion. For the shorter CORE passages, on average, the ASR scores underestimated and the traditional human scores overestimated the WCPM compared to the criterion score. These results could have implications in an applied setting where CBM-R scores are used to identify students at risk of poor reading outcomes.

For example, the inflated traditional human CBM-R scores could lead educators to underidentify at-risk students on universal screening assessments, as the student scores would be greater than what might be expected.

In addition, all 12 pairwise comparisons (3 grades × 4 passage lengths) between the human criterion and the ASR scores were statistically significantly different, and eight of the 12 pairwise comparisons between the human criterion and the traditional human scores were statistically significantly different. Thus, both the traditional human and ASR scores were different from the human criterion, though the ASR scores were more often so. This finding supports the notion that human scoring is generally more reliable than automated scoring (Williamson et al., 2012); however, nine of 12 pairwise *ES*s between the human criterion and the ASR scores (and 11 *ES*s between the two human scores) were below the standardized mean score difference threshold of .15 proposed by Williamson et al. These results support the use of ASR to score CBM-R assessments in schools, even when they are used for consequential assessment.

WCPM –Passage Length

In response to our second research question, we also examined differences in WCPM scores between traditional CBM-R passage lengths (about 250 words, read for 60 s) and shorter passages (25 to 85 words, read in their entirety). We found no statistically significant differences in the 27 pairwise comparisons (3 passage lengths comparisons × 3 scoring methods × 3 grades) of WCPM scores between traditional CBM-R passages and the shorter CORE passages (see Table 7).

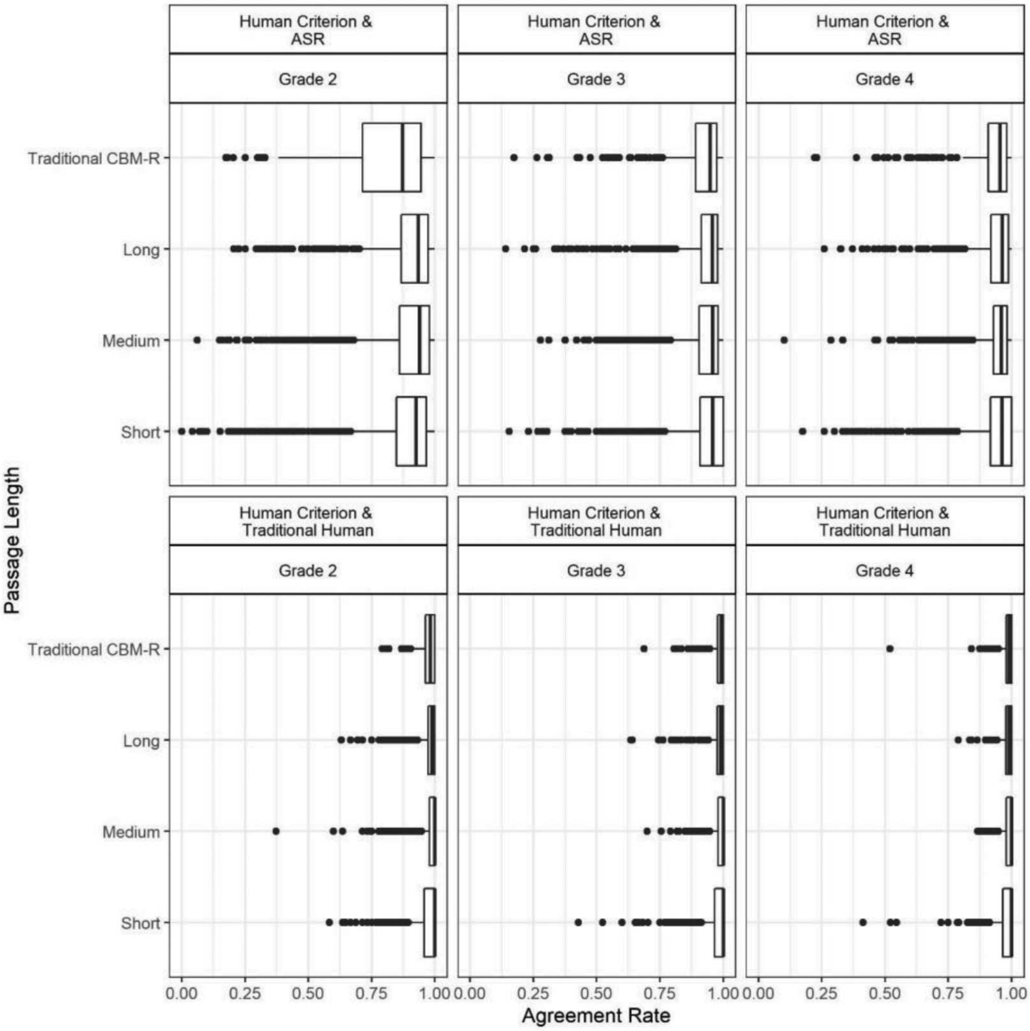

Figure 1. Box plot distributions of word score agreement rates for human criterion scoring and automatic speech recognition (ASR) for traditional human scoring by grade and passage length. CBM-R = curriculum-based measurement of oral reading fluency.

Thus, the scores for the shorter CORE passages were generally comparable to the scores for the traditional CBM-R passages (irrespective of scoring method), which provides evidence that the shorter passages can be used in schools in the administration of CBM-R. This finding has the potential to make a sizable contribution to the field under the idea that traditional CBM-R administration can be made even briefer (e.g., 30 s), saving time and resources. We speculate, however, that traditional practices are unlikely to change, given the convenience of the status quo 60 s assessment.

Nevertheless, research has shown that the use of shorter passages, in combination with advanced psychometric modeling, has the potential to reduce the standard error and increase the reliability of WCPM scores (Kara et al., 2020). A smaller standard error for CBM-R scores would provide educators with more accurate scores with which to monitor student CBM-R progress over time (Nese & Kamata, 2020), and thus increase the consequential validity of the decisions based on those scores.

Word Agreement

In response to our third research question, we found that the human-to-ASR degradation from the human-to-human score agreement was generally below the threshold of .10 recommended by Williamson et al. (2012; Table 8). In other words, the differences between human-to-human word score agreement and human-to-ASR word score agreement were generally appropriate for consequential assessment.

We found, on average, very high word score agreement rates between the human criterion and ASR scores (.81 to .94). If, according to Zechner et al. (2012), the "known" ASR word accuracy rate for students is 71–85%, then all but one of these average human-to-ASR agreement rates exceeded the highest end of that expectation.

Given these results, stakeholders and educators can weigh the balance between high accuracy of human assessors and the resource demands of one-to-one CBM-R administration. That is, although there may be some loss in accuracy of CBM-R word scores using ASR, gains stand to be made in terms of the costs in time and human resources of CBM-R administration. In addition, the loss in word accuracy is less evident when looking at WCPM scores, as demonstrated by the results of our second research question. That is, the ASR timings were more accurate than traditional human timings; so if used operationally, we speculate that the resulting WCPM scores could be equally or more reliable than traditional human scores because they reduce error associated with having humans accurately calculate a 60 s interval as they are actively engaged in administering a CBM-R in person. Even more, the trade-off between accuracy and resources may level in time as technology advances and ASR engines become more accurate. Although the highest "known" ASR word accuracy rate in 2012 was up to 85% (Zechner et al., 2012), that range may need to be adjusted with technological advances. The ASR applied in this study was developed in 2016, and we speculate a more current ASR may have better word accuracy, but this speculation needs to be examined in future research.

Limitations

A number of limitations in the present study should be noted and considered when interpreting results. To begin, although a large amount of data was collected for this study, some information was unavailable, including complete student demographic characteristics, which limits generalization of the findings to other populations. Also, one of the administration errors regularly associated with traditional human CBM-R scores (miscalculation of WCPM) was not present in this study, which likely influenced results. Despite the evidence of human errors in calculating WCPM in traditional administration (Reed et al., 2014; Reed & Sturges, 2013), our traditional human WCPM scores were not calculated by the human scorer. Rather these WCPM scores were calculated automatically by the computer using the reported words correct and time to read a passage. Thus, we speculate that the traditional human scores used here may have been more reliable than what would be typically expected in schools. Another possible confound was that the traditional human scorers in this study were well-trained, and many quite experienced in CBM-R administration, which may not be representative of real-world settings (Munir-McHill et al., 2012).

Another way in which our study's data collection differed from typical school-based practice involves participating students read aloud for 6 to 7 min in one continuous session. It is conceivable that student fatigue may have influenced the results, which may limit the extent to which the results can be generalized to school-based practice. An additional possible limitation involved the way in which we calculated the time it took for students to read the CORE measures under the human criterion scoring condition. Because the human criterion scores did not include a measure of time, we chose to use the ASR time when calculating the human criterion WCPM scores. We chose the ASR times under the assumption that they would be more accurate than the traditional human times, which consequently favored the ASR WCPM scores in our comparisons, and likely influenced results.

Finally, it was suspected that the use of the LMM might not have been the best choice for the outcome variables, because they took only positive values and were positively skewed. Therefore, we examined the distributions of residuals for the final models for both reading time recorded and WCPM and found that the residuals were also positively skewed for both recorded reading time and WCPM. However, the distribution of the residuals across the predicted outcome values did not display any nonrandom pattern for WCPM, but did display some mild nonrandom patterns for recorded reading time. Thus, we fitted two alternative models for recorded reading time: (a) LMM using logarithmic transformed outcome variables, and (b) log-linked Gamma generalized linear mixed-effect models (GLMM). These alternative models did not yield consistent improvements on the residual distributions, so we decided to retain the LMM for both recorded reading time and WCPM, but we acknowledge that other researchers may have decided differently.

Conclusions and Future Directions

In general, our findings are largely consistent with the small number of previous studies showing that ASR can provide reliable WCPM scores compared to expert human scorers (Bernstein et al., 2017; Bolaños et al., 2011; Zechner et al., 2009; Zechner et al., 2012), but also extend previous findings and address several gaps in the research. Ours is the first study to compare criterion WCPM scores obtained in a clinical setting using recordings to enable score verification, to both ASR and traditional human CBM-R scores consistent with those conducted in schools. These comparisons allow for the analysis of the potential utility of ASR compared to current school practices (as opposed to scores based on audio recordings), which we speculate is a more useful metric for educators, administrators, school district officials, and stakeholders. In addition to examining WCPM across scoring methods, we analyzed time and word accuracy agreement, which allow for a more nuanced examination of score differences besides aggregated WCPM passage scores. Last, we examined whether shorter passages can adequately replace traditional CBM-R passages, which might be an important part of reducing the large *SE* associated with traditional CBM-R fluency scores.

The results presented here provide preliminary evidence that ASR can be used in schools to score CBM-R for consequential assessment, based on criterion suggested by Williamson and colleagues (2012), and that shorter passages can be used in CBM-R assessment. A computerized CBM-R system that incorporates ASR and multiple shorter passages (combined with advanced psychometric modeling; Kara et al., 2020) has the potential to (a) reduce human administration errors by standardizing administration setting, delivery, and scoring; (b) reduce the time cost of CBM-R administration by allowing small-group or whole-classroom testing, (c) reduce the resource cost to train staff to administer and score the assessment; and (d) collect word-level time data that can be used in a model-based approach to scale WCPM scores and reduce the standard error of CBM-R measurement. Thus, this study is an important part of a larger effort to improve traditional CBM-R assessment to then improve systems used by educators to make data-based decisions, the results of which are the charge of future research.

References

Alonzo, J. (2016). The relation between Smarter Balanced and easyCBM mathematics and reading assessments. *Journal of School Administration Research and Development, 1,* 17–35.

Alonzo, J., Tindal, G., Ulmer, K., & Glasgow, A. (2006). *easyCBM© online progress monitoring assessment system*. Eugene, OR: Behavioral Research and Teaching. Retrieved from http://easyCBM.com

Anderson, D. (n.d.). esvis: Visualization and estimation of effect sizes (R package version 0.2.0) [Computer software]. Retrieved from https://github.com/DJAnderson07/esvis

Anderson, D., Alonzo, J., Tindal, G., Farley, D., Irvin, P. S., Lai, C. F., . . . Wray, K. A. (2014). *Technical manual: EasyCBM* (Technical Report No. 1408). Eugene, OR: Behavioral Research and Teaching.

Ardoin, S. P., & Christ, T. J. (2009). Curriculum-based measurement of oral reading: Standard errors associated with progress monitoring outcomes from DIBELS, AIMSweb, and an experimental passage set. *School Psychology Review, 38,* 266–283. http://dx.doi.org/10.1080/02796015.2009.12087837

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1–48. http://dx.doi.org/10.18637/jss.v067.i01

Bernstein, J., Cheng, J., Balogh, J., & Rosenfeld, E. (2017). Studies of a self-administered oral reading assessment. *Proceedings of the 7th ISCA Workshop on Speech and Language Technology in Education* (pp. 172–176). Stockholm, Sweden: ISCA. http://dx.doi.org/10.21437/SLaTE.2017-30

Bolaños, D., Cole, R. A., Ward, W., Borts, E., & Svirsky, E. (2011). FLORA: Fluent oral reading assessment of children's speech. [TSLP]. *ACM Transactions on Speech and Language Processing, 7,* 1–19. http://dx.doi.org/10.1145/1998384.1998390

Christ, T. J. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimating standard error of the slope to construct confidence intervals. *School Psychology Review, 35,* 128–133. http://dx.doi.org/10.1080/02796015.2006.12088006

Christ, T. J., & Coolong-Chaffin, M. (2007). Interpretations of curriculum-based measurement outcomes: Standard error and confidence intervals. *School Psychology Forum, 1,* 75–86.

Christ, T. J., & Silberglitt, B. (2007). Estimates of the standard error of measurement for curriculum-based measures of oral reading fluency. *School Psychology Review, 36,* 130–146. http://dx.doi.org/10.1080/02796015.2007.12087956

Christ, T. J., Zopluoglu, C., Long, J. D., & Monaghen, B. D. (2012). Curriculum-based measurement of oral reading: Quality of progress monitoring outcomes. *Exceptional Children, 78,* 356–373. http://dx.doi.org/10.1177/001440291207800306

Clarkson, P., & Rosenfeld, R. (1997, September). *Statistical language modeling using the CMU-Cambridge toolkit*. Speech presented at the Fifth European Conference on Speech Communication and Technology, Rhodes, Greece.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46. http://dx.doi.org/10.1177/001316446002000104

Cohen, J. (1988). *Statistical power analysis for the social sciences*. Hillsdale, NJ: Erlbaum.

Colón, E. P., & Kranzler, J. H. (2006). Effect of instructions on curriculum-based measurement of reading. *Journal of Psychoeducational Assessment, 24,* 318–328. http://dx.doi.org/10.1177/0734282906287830

Cummings, K. D., Biancarosa, G., Schaper, A., & Reed, D. K. (2014). Examiner error in curriculum-based measurement of oral reading. *Journal of School Psychology, 52,* 361–375. http://dx.doi.org/10.1016/j.jsp.2014.05.007

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 18,* 19–32.

Derr-Minneci, T. F., & Shapiro, E. S. (1992). Validating curriculum-based measurement in reading from a behavioral perspective. *School Psychology Quarterly, 7,* 2–16. http://dx.doi.org/10.1037/h0088244

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,* 239–256. http://dx.doi.org/10.1207/S1532799XSSR0503_3

Hoffman, A. R., Jenkins, J. E., & Dunlap, S. K. (2009). Using DIBELS: A survey of purposes and practices. *Reading Psychology, 30,* 1–16. http://dx.doi.org/10.1080/02702710802274820

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C. L., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology, 95,* 719–729. http://dx.doi.org/10.1037/0022-0663.95.4.719

Kara, Y., Kamata, A., Potgieter, C., & Nese, J. F. T. (2020). Estimating model-based oral reading fluency: A Bayesian approach. *Educational and Psychological Measurement, 80,* 847–869. http://dx.doi.org/10.1177/0013164419900208

Lenth, R. (2019). emmeans: Estimated marginal means, aka least-squares means (R package version 1.3.2.) [Computer software]. Retrieved from https://CRAN.R-project.org/package=emmeans

Müller, K. (2017). here: A simpler way to find your files (R package version 0.1.) [Computer software]. Retrieved from https://CRAN.R-project.org/package=here

Munir-McHill, S., Bousselot, T., Cummings, K. D., & Smith, J. L. M. (2012, February). *Profiles in school-level data-based decision making*. Paper presented at the National Association of School Psychologists 44th Annual Convention, Philadelphia, PA.

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: Reports of the subgroups* (NIH Publication No. 00–4754). Washington, DC: U.S. Government Printing Office.

Nese, J. F. T., & Kamata, A. (2020). Addressing the large standard error of traditional CBM-R: Estimating the conditional standard error of a model-based estimate of CBM-R. *Assessment for Effective Intervention.* Advance online publication. http://dx.doi.org/10.1177/1534508420937801

Nese, J. F. T., Park, B. J., Alonzo, J., & Tindal, G. (2011). Applied curriculum-based measurement as a predictor of high-stakes assessment: Implications for researchers and teachers. *The Elementary School Journal, 111,* 608–624. http://dx.doi.org/10.1086/659034

Pedersen, T. L. (2019). patchwork: The composer of plots (R package version 1.0.0.) [Computer software]. Retrieved from https://CRAN.R-project.org/package=patchwork

Pinnell, G. S., Pikulski, J. J., Wixson, K. K., Campbell, J. R., Gough, P. B., & Beatty, A. S. (1995). *Listening to children read aloud*. Washington, DC: Office of Educational Research and Improvement, U. S. Department of Education.

Poncy, B. C., Skinner, C. H., & Axtell, P. K. (2005). An investigation of the reliability and standard error of measurement of words read correctly per minute using curriculum-based measurement. *Journal of Psychoeducational Assessment, 23,* 326–338. http://dx.doi.org/10.1177/073428290502300403

Potgieter, C. J., Kamata, A., & Kara, Y. (2017). *An EM algorithm for estimating an oral reading speed and accuracy model*. Retrieved from https://arxiv.org/abs/1705.10446

R Core Team. (2019). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reed, D. K., Cummings, K. D., Schaper, A., & Biancarosa, G. (2014). Assessment fidelity in reading intervention research: A synthesis of the literature. *Review of Educational Research, 84,* 275–321. http://dx.doi.org/10.3102/0034654314522131

Reed, D. K., & Sturges, K. M. (2013). An examination of assessment fidelity in the administration and interpretation of reading tests. *Remedial and Special Education, 34,* 259–268. http://dx.doi.org/10.1177/0741932512464580

Robinson, D., & Hayes, A. (2018). broom: Convert statistical analysis objects into tidy tibbles (R package version 0.5.1.) [Computer software]. Retrieved from https://CRAN.R-project.org/package=broom

Saven, J. L., Tindal, G., Irvin, P. S., Farley, D., & Alonzo, J. (2014). *easyCBM norms 2014 edition* (Technical Report No. 1409). Eugene, OR: Behavioral Research and Teaching. Retrieved from https://files.eric.ed.gov/fulltext/ED547421.pdf

Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment, 24,* 19–35. http://dx.doi.org/10.1177/0734282905285237

Speece, D. L., Case, L. P., & Molloy, D. E. (2003). Responsiveness to general education instruction as the first gate to learning disabilities identification. *Learning Disabilities Research & Practice, 18,* 147–156. http://dx.doi.org/10.1111/1540-5826.00071

Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41,* 85–120. http://dx.doi.org/10.1177/00224669070410020401

Wickham, H. (2017). tidyverse: Easily install and load the 'tidyverse' (R package version 1.2.1.) [Computer software]. Retrieved from https://CRAN.R-project.org/package=tidyverse

Wickham, H., Hester, J., & Chang, W. (2018). devtools: Tools to make developing R packages easier (R package version 1.2.1.) [Computer software]. Retrieved from https://CRAN.R-project.org/package=devtools

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31,* 2–13. http://dx.doi.org/10.1111/j.1745-3992.2011.00223.x

Wood, D. E. (2006). Modeling the relationship between oral reading fluency and performance on a statewide reading test. *Educational Assessment, 11,* 85–104. http://dx.doi.org/10.1207/s15326977ea1102_1

Zechner, K., Evanini, K., & Laitusis, C. (2012). Using automatic speech recognition to assess the reading proficiency of a diverse sample of middle school students. *Third Workshop on Child, Computer and Interaction* (pp. 45–52). Portland, OR: ISCA. Retrieved from https://www.isca-speech.org/archive/wocci_2012/papers/wc12_045.pdf

Zechner, K., Sabatini, J., & Chen, L. (2009). Automatic scoring of children's read-aloud text passages and word lists. *Proceedings of the NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications* (pp 10–18). Boulder, CO: Association for Computational Linguistics.

Received January 23, 2020
Revision received August 27, 2020
Accepted September 19, 2020 ■