A paper presented at the ***California Educational Research Association*** (CERA) 96th Annual Conference, November 2017.

Oral Reading Assessment: Leveled Fluency, Self-Administered and Automatically Scored

Jared Bernstein*†, John Sabatini‡, Jennifer Balogh*, Jian Cheng*
* Analytic Measures Inc., †Stanford University, ‡Educational Testing Service

In Grades 1 through 5, literacy standards often require that students read with sufficient accuracy and fluency to support comprehension. These skills can be measured with oral reading fluency (ORF) assessments. However, ORF assessments are very time-intensive because they require teachers to administer and score each test individually for every student. Additionally, scores can be subject to several sources of uncontrolled, irrelevant variance. To improve ORF assessments, a prototype mobile assessment (Moby.Read) was designed and implemented as a fully automated, self-administered, tablet-based reading test that uses speech recognition and speech processing technologies to measure accuracy and fluency of oral reading. We report on the design, development, and evaluation of this prototype, and present data that address several research questions including: how usable is the app? and how accurate are the scores? The impact of the technology on reading assessment is discussed.

Background

Reading is a skill that is critical for success in school and in life. Because of the importance of reading, educational standards specify foundational skills of reading that must be met in each grade, and reading fluency has recently been emphasized as one of these skills (NGACBP & CCSSO, 2010).

Theory of change and empirical support

An automated ORF assessment is an enhanced and efficient form of curriculum-based measure (CBM). CBMs are classroom-based assessments that identify which students need additional support to reach grade-level proficiency. CBMs are often used as screening or progress monitoring tools within a Response to Intervention (RTI) model. CBMs should be relatively easy to administer, brief in duration, and provide an achievement score in a domain such as reading, writing, or mathematics (Deno, 1985). Oral reading fluency is a CBM that generally tracks (covaries with) reading comprehension for most children below grade 5. In an ORF measure, students read a word list or a passage aloud, and their speed and accuracy is recorded. The ORF test format is usually a one- or two-minute reading of a grade-level passage, to produce a WCPM (words correct per minute) score (Deno, 1985; Wayman et al., 2007). Popular ORF instruments include DIBELS (Good, Kaminski, Cummings, Dufour-Martel, Peterson, Powell-Smith, & Wallin, 2011). and AIMSweb. (Pearson, 2012).

Research results indicate that passage ORF is a stronger and more efficient indicator of reading comprehension than word-list reading rate (Jenkins et al., 2003). Studies also show the utility of CBMs to predict student performance on high-stakes state reading assessments (Crawford, Tindal, & Stieber, 2001).
A theory consistent with these empirical results is that ORF WCPM may reflect the automatization of essential components of reading comprehension (including schematic reasoning and syntactic parsing). In this view, ORF is not merely a convenient marker, but is actually a direct indicator of overall reading competence. ORF continues as a practical measurement tool, but also has theoretical value in understanding reading processes and development (Kuhn et al., 2010; Eason et al., 2013). For example, reading fluency is one of five component skills that foster the development of good reading comprehension outcomes (NICHD, 2000). Unlike silent reading, oral reading makes the cognitive process of reading more observable, which creates an evidence trail to support inferences about which challenges young readers may be experiencing in acquiring efficient decoding, word recognition, continuous text fluency (expressive prosody) and even some aspects of comprehension. For some of these reasons, the Common Core State Standards for Grades 1 through 5 (http://www.corestandards.org/ELA-Literacy/RF/) call out reading fluency as a foundational skill.

Although ORF assessments are theoretically beneficial, the efficiency, reliability and utility of traditional ORF measurement (as practiced in real schools) has been challenged in recent research with respect to:

(a) inefficient use of teacher time with traditional ORF measures (Cummings et al., 2014);
(b) human scoring error from training deficits (Cummings et al., 2014);
(c) emphasis on speed instead of expression of meaning (Schwanenflugel & Benjamin, 2012);
(d) teacher uncertainty about how to use ORF scores (Deeney & Shim, 2016); and
(e) score instability from passage & individual factors (Ardoin et al. 2013; Francis et al., 2008).

The prototype design addresses all five of these problems – self-administration and automatic scoring radically reduce teacher time; automatic scoring is more accurate and consistent than the scoring available from an average, busy teacher in need of in-service training. With accurate automatic analyses of expression, an automated reading assessment should emphasize good oral reading as opposed to hurried oral reading; Speech-based analytics can enhance the quality of inferences one can make about the nature of student weaknesses and the promising remediation paths to help students towards proficiency (Deeney, 2010); and new empirical techniques applied to equating passages across test sessions will reduce individual error.

Rasinski (2004) offers an operational definition of fluency as the "ability to develop control over surface-level text processing so that [the student] can focus on understanding the deeper levels of meaning embedded in the text." Rasinski underscores several important points in this definition: first is control over basic text processing. If a student has not cracked the code that pairs written symbols and spoken words, then oral reading cannot take place. The second point is that the purpose of decoding the text is to unlock its meaning.

Ideas about the intricate interaction between the mental processes required to process written text and comprehension were inspired by LaBerge and Samuels' seminal work on a theory of automatic processing in reading. LaBerge and Samuels (1974) posit that there are limited mental resources available for processing information. When processing text is slow and labored, mental energy is predominantly spent on the mechanics of translating written symbols to sounds, and therefore little or no mental energy is left for understanding and integration of the text with the reader's prior knowledge and worldview.

However, when processing text is automatic for the reader, little mental energy is consumed on decoding letter-to-sound patterns and more energy is available to grasp and process the meaning of the words in a text. Highly proficient readers integrate words and phrases they are currently reading with surrounding sentences and paragraphs. In this way, skilled readers are able to comprehend the text at a high-level as they read it and can generally appreciate whether or not upcoming words fit within the context of what is being read. At this level of skill, the reader understands what words should be emphasized and how words fit together into larger units because the reader is comprehending in real-time. That real-time understanding can often be heard in the expressiveness of an oral reading.

Building on these concepts, Rasinski describes three important dimensions to fluent reading: accuracy in word decoding, automatic processing (reflected in accurate reading rate), and prosodic reading (reading with appropriate phrasing, pausing, emphasis and expression). These three concepts are consistent with other descriptions of reading fluency including a report by the National Reading Panel, which states that "fluent readers can read text with speed, accuracy, and proper expression" (NICHHD, 2000, p 3-1).

Reading Fluency Assessments

Assessments have been created to measure these three dimensions of Oral Reading Fluency (ORF). The most common and basic method of assessing oral reading involves a read-aloud procedure in which a student reads passages out loud and a teacher times the student while marking specific reading errors. The teacher may also provide a subjective rating of the student's prosody. ORF assessments have enjoyed widespread acceptance in the

classroom and are used by teachers for many purposes including placing students in reading groups, identifying students who need intervention, guiding individual instruction, benchmarking performance over the year, and monitoring progress (Deeney & Shim, 2016).

However, traditional ORF assessments present several problems. For one, they require an excessive amount of teacher time. This time is taken up with in-service teacher training devoted to instructing teachers on how to administer the assessments, how to mark reading errors according to the rules of the specific assessment, and (for some tests) how to select appropriate leveled passages according to detailed procedures. Then there is the ongoing administration of the assessments individually to each student, followed by tallying errors and calculating summary scores. Finally, more teacher time is spent entering results by hand, usually to a website that tracks scores.

Consider the total time spent on ORF administration alone. If there are about 20 million students enrolled in U.S. public elementary schools and 20% of these students are given ORF assessments three times a year and each assessment takes 10 minutes, then the total time spent per year on administering ORF assessments is 2 million hours of teacher time. If the information from these assessments could be gathered from students' self-administered assessments, then some of those 2 million hours of time could be redirected toward instruction. Further, assessments could be administered more frequently with much less teacher effort.

Another problem with ORF assessments mentioned by districts is that many teachers administer or score ORF assessments incorrectly. Accuracy and consistency within and across test sessions is important for linking results with appropriate instruction and accurately monitoring progress at different times and between grades when different teachers are assessing students. For example, a child may be assessed in first grade by a teacher who is very lenient. The next year, the child may be assessed by a second grade teacher who is more strict, resulting in a lower reading level, even though the child's actual level is not significantly lower. This inconsistency across years introduces confusion for a parent who might try to decide whether or not to invest in private tutoring.

The Moby.Read prototype was designed to address these problems with ORF assessment. The vision is an assessment that students can self-administer to simplify scheduling and save teacher time. As implemented, the prototype uses speech recognition and speech processing technologies to score the assessment automatically and consistently across students and across sessions, which also reduces burden on classroom teachers. Note however that teachers can still listen to recordings of the students' readings and retellings and their answers to comprehension questions. Even in this prototype, recordings are available for review on the mobile device.

Description of the Prototype Assessment (Moby.Read)

Moby.Read is a self-administered and automatically scored ORF assessment. The prototype has been implemented as a mobile app that runs on Apple iPads in the iOS operating system. The iOS platform was chosen because it supports rapid prototyping with many convenient audio and touchscreen capabilities built in. The prototype app is designed to be used by students in Grades 1 through 5 and exploits speech recognition and speech processing technologies to detect, analyze and score oral reading and other spoken responses, such as retellings and answers to comprehension questions. Because the app successfully processes the child's speech (even with significant background noise), it moves through each step of the assessment protocol automatically at the student's pace. Once the student has completed the assessment, the teacher can immediately access a score report and analyses of the reading performances.

Figure 1 shows a schematic of the architecture of Moby.Read and how the different components interact with one another.

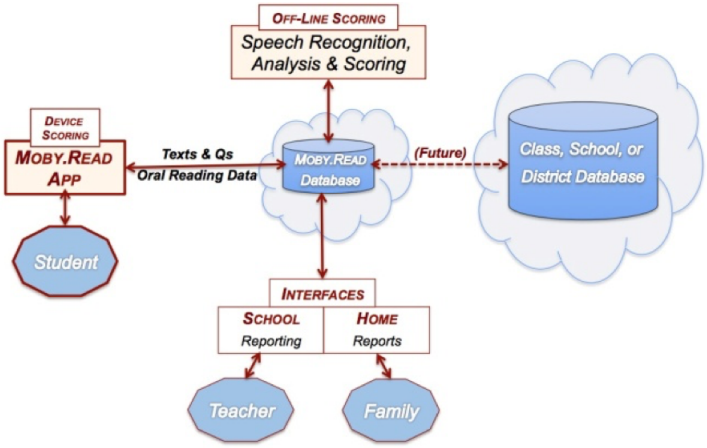

Figure 1. Schematic of Moby.Read architecture.

Once a test session has been completed, raw and analyzed information is sent securely to a cloud database. Additional analyses are possible off-line with more computing power. Once the database has been populated, a user interface from a secure website can be used to access the scores and recordings. Eventually, the assessment will be integrated with existing student information systems.

Student Experience

After launching the assessment, the student enters a PIN. If this is the first time the student is taking the test, an instructional video is presented that explains what to do and shows examples of a 4th grade boy performing these tasks. First, students are asked to read a word list. Next, the assessment presents a short introduction to a practice passage. Then the assessment presents the text to be read accompanied by a picture. As the student reads the passage out loud, the student's voice is digitized and analyzed. The assessment detects when the student has finished reading and then prompts the student to retell the passage in his or her own words. Next the student is asked a comprehension question about the passage. The student answers this question aloud with a word or short phrase. Then the assessment presents another comprehension question. After the student has completed the tasks for the practice passage, the assessment presents another passage with the same accompanying steps: an introduction, the text (with picture) for reading aloud, a prompt to retell the passage, and two comprehension questions.

The prototype version of the app presented three passages on grade level as is common in many ORF assessments. A future version may include an adaptive feature that adjusts the difficulty of the passages based on the student's reading performance and comprehension (combined with other criteria such as performance on a word list, teacher input, etc.).

Teacher Experience

Teachers, reading specialists, and administrators are the intended score users. To access the information available to score users, a teacher PIN is entered on the mobile device. The first page displays Student IDs, student names, and basic ORF measures including accurate reading rate (WCPM) and comprehension. From this page, score users can view a graph showing performance over time. Also, the score user can go to a different page to view the text that was presented to the student and listen to a recording of the student's reading, the student's retelling and the student's answers to two comprehension questions.

In the experiments reported below, the score user's information is available on the device that the student used to take the test. Eventually, the information should also be available from a secure website so that student assessments from any device will be accessible in one location.

Design Process for the Moby.Read Assessment

The development of the Moby.Read app underwent several iterations based on expert and user feedback. First, existing reading fluency assessments were surveyed and reviewed for flow, content and reporting structure to understand what was currently being used and what educators' expectations were.

Next, a storyboard was created for the prototype. The storyboard was a simple, low-fidelity slide set that showed screenshots of the different steps of the assessment. This storyboard served as a catalyst for discussion of the prototype and to get feedback from several reading experts and teachers. The feedback from this stage led to several changes in the design. For example, a confirmation step was added to ensure that the student about to take the test was the intended person. Actual usage patterns confirmed that this step caught and corrected careless ID entry errors. Another example of a change from the expert review was the addition of a practice passage. Again, subsequent usage showed that children benefited from practicing before being launched into the actual assessment. For example, one child did not speak immediately but then thinking back to the video eventually started interacting with the app. The practice passage allowed him to get used to talking to the app out loud without being penalized.

In the next step, a working prototype was created. The prototype was presented to a handful of students and stakeholders including reading specialists, teachers, principals and a technology coach. Again, feedback from this stage instigated another round of changes. For example, the scoring graph was modified to encompass a broader range of reading rates. It was discovered that some students were not certain what to do for the retell step in the assessment, so some of the prompts and questions were rewritten and rerecorded to more clearly describe what to do. Several pictures were redrawn so that they did not give away story elements that were the focus of comprehension questions. Also, font colors were adjusted to make it easier to see where to begin reading and to provide more coherence when the story was being read back to the student.

Finally, a larger-scale usability test was run in several classrooms. Details of this study are described below.

Passage Development, Leveling, and Equilibrating

An initial set of passages was written by a reading specialist. A second set was written by a professional content developer. The second set was reviewed by the reading specialist and revised if necessary. All passages were then reviewed by a panel of assessment experts.

The Council of Chief State School Officers (CCSSO&NGA, 2012) associated with the Common Core State Standards (CCSS) describes a three-part model for judging text complexity. Part I is a quantitative step of using one of the following text complexity tools to level the text: ATOS, Degrees of Reading Power, Flesch-Kincaid, Lexiles, Reading Maturity, or SourceRater (see CCSSO&NGA, 2012). Part II is a qualitative expert evaluation of text that considers four dimensions: structure, language conventionality and clarity, knowledge demands, and levels of meaning (literary) or purpose (informational). Part III of the model is a consideration of the reader and the task. This part of the model calls on professional educators to use their judgment in selecting texts for a specific task or matching a certain group of students with texts (for example students who might only do well with high interest texts).

Adhering to this method, all passages were first analyzed according to the Flesch-Kincaid (Flesch, 1948) readability formula. Independently, a reading specialist reviewed the passages according to the four dimensions above and subjectively assigned them a level. Finally, all passages were reviewed to ensure that they would be appropriate as a reading task in the context of a reading assessment.

In many cases, there was a discrepancy between the level estimated by the Flesch-Kincaid readability formula and the more subjective level assigned by the reading specialist. For a difference of one or two levels, the reading specialist considered whether or not the level from the readability formulas was reasonable. If so, the reading specialist adjusted the subjective level. If the reading specialist felt the level should not be changed, the level from the reading specialist was retained. For passages with more than two levels of differences, passages were reworked so that the objective and subjective levels were reconciled.

Passages were then piloted with student readers. Through more piloting in the future, sufficient data will be collected for each passage such that difficulty can be estimated accurately. Then through statistical modeling, the measured scores will be adjusted such that the measurements for each passage in a given level will produce equilibrated scores.

Usability and Accuracy

To understand the accuracy and usability of the assessment, several research questions were addressed.

Research Questions

A. Can students in grades 2, 3, and 4 use the assessment independently?
B. Do students in the pilot study have a good experience with the assessment? Why or why not?
C. Do teachers find the assessment useful, intuitive and/or convenient?
D. Are the scores from the assessment similar to those provided by human scorers?
E. Does the assessment yield scores similar to those from traditional human-administered ORF tasks?

To address these research questions, two studies were conducted. In the first study (Study 1), classrooms of students in grades 2, 3, and 4 were administered the assessment at school. Facilitators observed whether or not each student could complete the assessment independently (to address Question A). A usability questionnaire at the end of the assessment and a short debriefing interview with a subset of students provided more detailed usability data (to address Question B). Also, a usability questionnaire was presented to the classroom teachers to gather teacher feedback about the assessment's usefulness, ease of use, and convenience (to answer Question C). To address Question D, the prototype's automatically generated scores for each student were collected. Separately, human raters listened to the recordings of the student reading performances. Then the device-generated scores were compared to the human scores. Finally, for Question E, a separate study was conducted (Study 2) in which students took the automated assessment and also the ORF section of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) *Next* assessment (Good et al., 2011), a widely used human-administered and human-scored ORF assessment (Study 2a) or the running record assessment of the Teacher's College Reading and Writing Project (TCRWP, 2014).. Scores from the Moby.Read assessment and the other, more traditional assessments were then compared.

Study 1

Method

Participants. Participants in Study 1 were 99 school-aged children from four different elementary schools: Roosevelt Public School in Roosevelt, New Jersey, Trenton Catholic Academy in Hamilton, New Jersey, Oak Knoll Elementary in Menlo Park, California and Fox Elementary in Belmont, California. The female to male ratio was 47:52. Ages ranged from 7 to 10 with an average age of 8. Students were enrolled in 2nd Grade (29%), 3rd Grade (40%) and 4th

Grade (31%). Participant ethnic backgrounds were (using classifications set forth by the US Census), 51% of the students were white, 19% were African American, 4% were Asian, and 25% were Hispanic or Latino.

Four teachers provided usability feedback by filling in the teacher questionnaire. The teachers were a 2nd and 3rd grade teacher from Roosevelt Public School in NJ, a 3rd grade teacher from Oak Knoll Elementary School in CA, and a 4th grade teacher from Trenton Catholic Academy in NJ.

Assessment. Three different forms were used in Study 1, depending on the student's grade. The only difference between the forms were the three passages presented to the students, which were different depending on the student's grade and were leveled to this grade based on Flesch-Kincaid readability levels. Also, the accompanying introductions and comprehension questions for the passages were different and depended on the passage.

At the end of the test was a usability survey. The usability prompt was embedded in the app itself for consistency. At the conclusion of the reading tasks, students were presented with a page that asked how easy the app was to use. Then the student was prompted to pick one of four emoji faces. The faces were revealed one at a time with an oral description of what each meant, which was read to the student as the face appeared on the screen:

"Totally confused" "Not sure" "I knew what to do most of the time." "Easy. I always knew what to do."

Figure 2. Emojis used in student usability ratings.

The app registered student touches after all faces had been presented and described. *Ease of use* was the focus instead of liking because in pilot trials, it was discovered that children did not differentiate "*liking the passages*" from "*liking the app.*"

Separately, the teachers were presented with their own Teacher Questionnaire. The questionnaire was a Likert-style survey with 14 items that probed seven different dimensions of the app: general liking, intuitiveness for kids, appropriateness of content, usefulness of automatic administration, usefulness of scoring, preference for voice, and liking of the visual appearance. The Likert questionnaire is presented in Appendix A.

Procedure. For Study 1, two facilitators assisted with test administration: one in New Jersey and one in California. Both facilitators were assessment professionals. The experimental sessions with student-participants were conducted during the normal course of a school day at the participant's elementary school. Before the sessions started in a class, the classroom teacher was encouraged to experience the student-assessment path in the app so that the teacher would see and hear what the students experienced in the experimental assessments.

In preparation for the experimental sessions, the facilitators set up two chairs about eight feet apart in a quiet area of the room or just outside the classroom. Teachers were encouraged to participate and observe students as they interacted with the app. Before the assessment, each student was fitted with a set of GearHead headphones with an inline microphone (the microphone was incorporated into the wire). This microphone headset was chosen for two reasons: first, the ear pieces could be wiped down (with disinfectant) easily between administrations (as opposed to headsets with foam cushions), and second, the microphone was inconspicuous to the students. (It was found in pilot trials that children often played with boom microphones). Facilitators were present to help with technical

problems, but they did not help students take the Moby.Read assessment. If a student asked a question during the assessment, the facilitator encouraged the student to keep trying with the app. After each session, the microphone headset was wiped down.

After students were run in a given session, and at a time convenient for the teacher, the teacher was offered the iPad and an ID for the Teacher Pages for the class. Teachers viewed score reports and were able to play back student reading performances. Then the facilitator presented the teacher with the Teacher Questionnaire.

Results

Task completion. First, task completion was analyzed. Of the 99 students who attempted an assessment with the Moby.Read app, 95% were able to go through the app and provide responses that were scored and made available to the teacher. Of the five students who were not able to provide data, two encountered technical problems in which the head phones were not plugged in and were not able to detect audio and three spoke too softly or read part of the passage silently.

Student usability. Next, results of the student usability survey were analyzed. In the survey, students were asked how easy the app was to use. Responses were coded in the following way:

1 = totally confused face
2 = not sure
3 = I knew what to do most of the time
4 = easy, I always knew what to do

Thus, a higher score indicated a higher degree of usability. Scores ranged from 2 to 4. The average score was 3.4 indicating a high level of usability. Figure 3 shows a histogram of the student usability results.

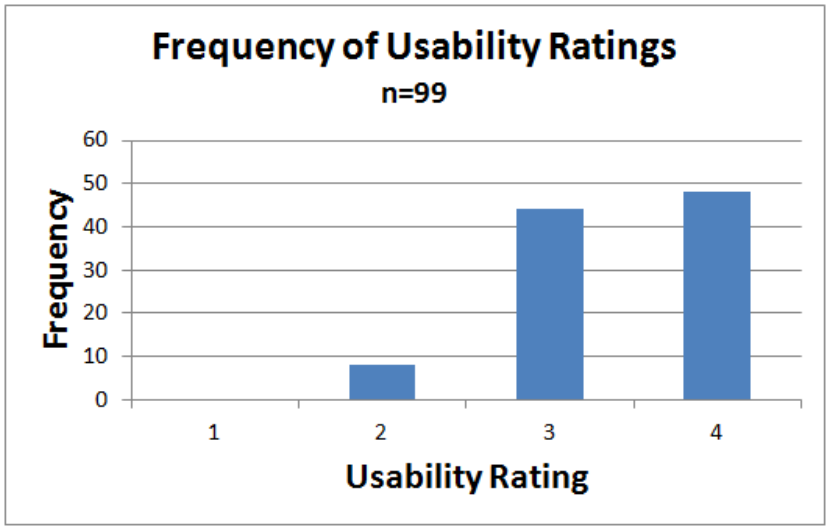

Figure 3. Histogram showing the distribution of student usability rating. 1 indicates a low usability rating and 4 a high rating.

As Figure 3 shows, no students were totally confused; only a few were unsure of what to do; and a plurality of students always knew what to do.

Teacher Usability. Responses to the Teacher Questionnaire were also analyzed. Responses were coded on a seven-point scale:
1 = strongly disagree
2 = disagree
3 = slightly disagree
4 = neutral
5 = slightly agree
6 = agree
7 = strongly agree

For negatively worded statements such as "I found Moby.Read frustrating to use," the scale was flipped such that a response of 2 was coded as 6. Table 1 presents the results of the Teacher Questionnaire.

Table 1. Results of the Teacher Questionnaire

Dimension	*Average Score (7 max)*
Like the App Overall	6.6
Intuitive for Kids	7.0
Appropriate Content	7.0
Usefulness of Automatic Administration	4.4
Usefulness of Scoring	5.2
Preference for Voice	5.8
Like the Visual Appearance	6.9
OVERALL Average Rating	6.1

Machine Scores Versus Human Scores. The Moby.Read assessment generated machine scores for each of the 94 students who completed the assessment. These scores were reported as words correct per minute (WCPM), a common metric in ORF assessments.

Separately, three human raters analyzed the reading performances of the students (the performances were automatically recorded by the Moby.Read assessment). The human raters computed WCPM by hand. Each recording was analyzed by at least two raters. Analysis of the human scores indicated that the inter-rater reliability of the human raters was extremely high at r = 0.99.

The correlation between human scores and automatic on-device Moby.Read scores was r = 0.96, which suggests that Moby's recognition and scoring closely match human scores. Moby.Read responses are also uploaded to AMI's servers. Server-based scoring featured more elaborate acoustic and language models. Server-based machine scores correlated with human scores with r = 0.987.

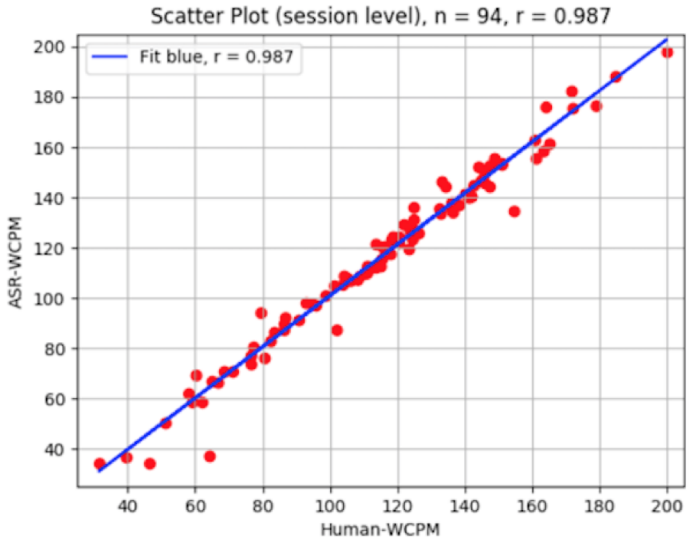

Figure 4. Session-level scatter of median WCPM; server-based scores vs. Human scores (r = 0.987).

The human WCPM in Figure 4 is a session-level value that is derived by averaging same-passage WCPM values from two raters and then using the median value per session.

Discussion

Task completion with the Moby.Read assessment was very high at 95%. The pilot study suggests that for initial use, the assessment may require some supervision to ensure students who do encounter technical difficulties (such as problems with a headset) are able to troubleshoot these issues. Also, it may take a small percentage of students who are shy about reading out loud some practice reading into the tablet. It is expected that task completion will continue to improve as students become more familiar with the assessment throughout the school year. Longer studies designed to monitor more extended use across a school year will be planned to identify changes that will promote improvements in task completion with practice. Redesigning the initial instructional video may help.

From a subjective standpoint, students seemed to know how to use the assessment independently as evidenced by the subjective scores on the student usability survey at the end of the assessment. Most students selected the choice of "always knew what to do," with only very few selecting "unsure."

For teacher usability, the results are promising. Overall, the assessment was rated very highly on average (6.1 out of 7). The lowest subscores warrant some discussion, however. The lowest subscores were for usefulness of automatic administration and usefulness of scoring, which the developers considered important advantages of this assessment technology. Delving deeper into these two areas, it was noted that some teachers were extremely enthusiastic about saving time, while others were neutral or slightly negative -- not because they did not appreciate the convenience, but because they felt that they would rather do the administrations themselves. Given that reading is such an important part of the curriculum, especially in the earlier grades, this tendency for some teachers to want to invest more time in their students' reading assessment is not unreasonable. This feedback has been useful for the development of Moby.Read and will inspire additional features such as including the option of hand-scoring the assessment.

Finally, with regard to the accuracy the correlation between human scores and automatic on-device Moby.Read scores was r = 0.987, which suggests that Moby's recognition and scoring closely match human scores.

In Study 1, the accuracy of scores was based on readings of passages presented in the Moby.Read test. The results demonstrate that the automatic scoring of the prototype is highly consistent with human scoring of the same reading performances. The next study extends this research by examining the validity of the scores. Scores from the prototype are purported to measure the construct of oral reading fluency. If these scores correlate highly with scores from other traditionally-used assessments that also claim to measure the same construct, oral reading fluency, then the results will provide evidence of construct validity.

Study 2

The goal of Study 2 was to address the final Research Question: Does the assessment yield scores similar to those from a traditional human-administered ORF task? Different kinds of scores are typically reported in different types of ORF assessments. One type emphasizes accurate reading rate with WCPM being the main reported score. A commonly-used assessment of this type is DIBELS NEXT. Another kind of score is the student's reading level. For assessments of this type, levels often range from A to Z with A being the easiest, and each level gradually increasing in difficulty as the alphabet progresses. A commonly-used assessment of this type is the running record of the Teacher's College Reading and Writing Project. Study 2 was conducted in two parts. In Study 2a, scores of accurate reading rate (WCPM) were compared with DIBELS NEXT. In Study 2b, Moby.Read scores were compared with reading levels from the Teacher's College assessment. The results from Study 2b were preliminary since the adaptive feature of Moby.Read was not yet implemented at the time of the study.

Study 2a

Method

Participants. Twenty students participated in Study 2a. Students were from Oak Knoll Elementary in Menlo Park, CA. Of the participants, 9 were female and 11 were male. Seven were in 2nd Grade; 6 were in 3rd Grade; and 7 were in 4th Grade.

Procedure. Students were administered both a Moby.Read assessment and a DIBELS NEXT assessment. For half the participants, Moby.Read was administered first, and for the other half DIBELS was administered first. Sessions were held at one private residence with the child's guardian close by, but not interfering with the session. The administrator was an assessment professional with experience using the DIBELS framework. For the Moby.Read assessments each child was fitted with a GearHead microphone headset.

For the DIBELS assessment, students were administered the fall benchmark test form, which consisted of three grade-leveled passages of about 250 words in length. The administrator followed the administration and scoring procedures described in the test's official documentation (Good, Kaminski & Cummings, 2011). When the student completed the session, the student was given a $15 gift card to a local toy store as remuneration. After the session, the administrator used the scoring sheets to calculate Words Correct per Minute (WCPM).

Results

Median WCPM scores are the reported scores for DIBELS, so the median DIBELS scores were compared to the median WCPM scores from Moby.Read for the same students. The correlation between the DIBELS WCPMs and Moby.Read WCPMs was $r = 0.88$. Figure 5 shows a scatterplot of the scores.

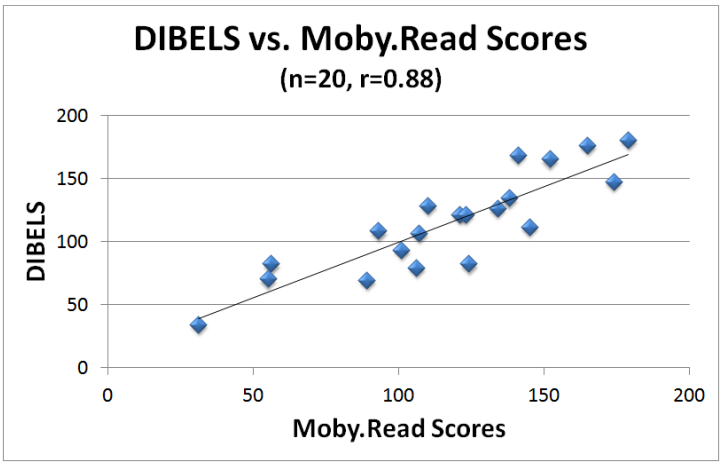

Figure 5. Scatterplot of Moby.Read WCPM scores and DIBELS WCPM scores.

Discussion

Results from Study 2a provide evidence that the Moby.Read assessment does yield scores that are similar to DIBELS, a traditional, human-administered and scored ORF assessment. The correlation between DIBELS WCPM scores and Moby.Read WCPM scores was 0.88. These results should be considered in light of several reliability measures. Published studies researching the DIBELS ORF assessment report a test-retest reliability of 0.82 and an inter-rater reliability of 0.85 (Goffreda & DiPerna, 2010). The reliability of a measurement instrument limits the strength of the correlation between that instrument and others measuring the same construct. Therefore, the correlation with Moby.Read is at the ceiling of what would be expected given the reliability of the DIBELS assessment.

Study 2b

The two assessments compared in Study 2b were Moby.Read and the running record assessment from the Teacher's College Reading and Writing Project.

Method

Participants. Twenty students participated in Study 2b. Students were from Martinez Unified School District. Not all students completed both tests at the time of this writing; therefore data were analyzed for 17 students.

Procedure. Students were administered a modified version of the Moby.Read assessment. At the time of the study, the adaptive feature in which the assessment automatically selected leveled passages was not yet implemented. Therefore, facilitators had students read several reading lists to determine an initial reading level and then presented the student with the appropriate test form in Moby.Read. The student took the test form in Moby.Read, which automatically scored the student's performance. Then, the facilitator presented easier forms or harder forms based on the assessment's automatic accurate-reading-rate score.

Separately, the participants were also administered a running record assessment from the Teacher's College Reading and Writing Project. The administrator was experienced at administering and scoring the assessment and followed the administration and scoring procedures described in the test's official documentation (TCRWP, 2014).

Sessions were held at the school in a real testing situation.

Results

Reading level is the reported score for the Teacher's College assessment. Final reading levels were converted to numbers where A = 1, B = 2, etc. Then these number were correlated with average WCPM scores from Moby.Read for the same students. The correlation between the reading levels and Moby.Read WCPMs was r = 0.94. Figure 6 shows a scatterplot of the scores.

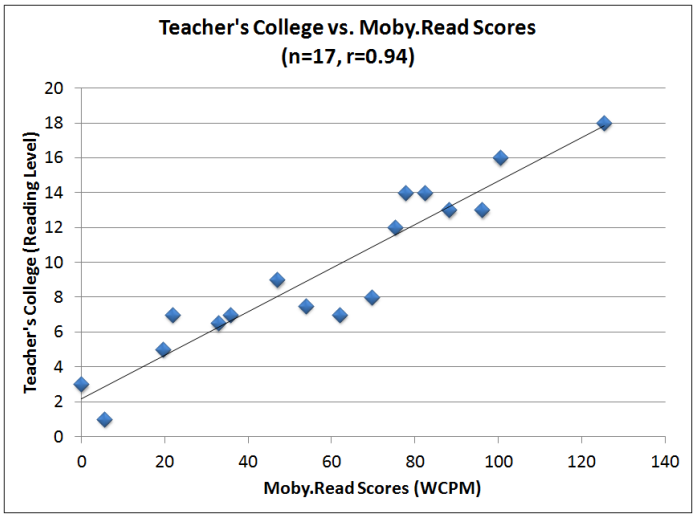

Figure 6. Scatterplot of Moby.Read WCPM scores and Teacher's College running record assessment reading levels.

Discussion

The results from Study 2b are preliminary given that the adaptive feature was not implemented at the time of the study. However, these results show a high correlation (r = 0.94) between Moby.Read's accurate reading rate scores and reading levels with the Teacher's College running record assessment. A follow-up study designed to corroborate this finding needs to be planned once an adaptive logic is implemented in the assessment app.

Conclusion

This paper described the design and development of a self-administered and automatically scored oral reading assessment called Moby.Read. The assessment is designed to help teachers assess a student's ORF ability more efficiently and more consistently than current hand-scored ORF assessments. Self-administration and automatic scoring will help teachers evaluate reading fluency more efficiently than current assessments while avoiding some inconsistencies.

The design of the prototype assessment has considered input from stakeholders during each step of the implementation process. The design has encompassed a rigorous method of statistical modeling to ensure that passages are appropriately leveled and equilibrated within level. The development integrates speech recognition and speech processing technologies to enhance the student's experience of a private, self-administered assessment.

This research has probed the usability and accuracy of the Moby.Read scores and has produced promising results. Five research questions were posed and were answered from the research:

A. Can students in grades 2, 3, and 4 use the assessment independently? **Yes, 95%**
B. Do students in the pilot study have a good experience with the assessment? **Yes (3.4 out of 4)**
   Why or why not? **(Students like technology and privacy)**
C. Do teachers find the assessment useful, intuitive and/or convenient? **Yes (6.1 out of 7)**
D. Are the scores from the assessment similar to those provided by human scorers? **Yes (r = 0.987)**
E. Does the assessment yield scores similar to those from traditional human-administered ORF tasks?
   Yes (r = 0.88 and r = 0.94)

The findings from Studies 1 and 2 showed that task completion rates are high (95%) and will most likely improve with practice, student usability is high (3.4 out of 4), teacher usability is high (6.1 out of 7), and accuracy is high (r = 0.987) for machine scores versus human scores, r = 0.88 for Moby.Read scores versus DIBELS scores, and 0.94 for a preliminary analysis of the automated scores versus human-rendered Teacher's College running-record scores).

Implication for Reading Assessments and Reading Instruction

The promising results from these studies suggest that a self-administered and automated assessment of oral reading fluency can produce scores that are accurate and can save teachers millions of hours of time that might be better spent on instruction. Further, the prototype assessment produces scores that are consistent, avoiding some potential problems introduced by irrelevant variance from uncontrolled sources.

The vision is that an assessment like Moby.Read will eventually extend beyond simply automating current ORF-style assessments and will provide more in-depth analyses in reading ability, beyond what human-administered and scored assessments can provide. With advancements in technology, more diagnostic information is possible since speech processing can track sub-passage timing, rate, and prosodic features, quicker and in more detail than a human scorer is able to. These possibilities are on the horizon as recent developments in spoken language processing open a new era of oral reading fluency assessment.

Acknowledgments

The research described here was supported by the Institute of Education Sciences, U.S. Department of Education, through the Small Business Innovation Research (SBIR) program contract ED-IES-16-C-0004 to Analytic Measures Inc. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

The authors acknowledge gratitude for the collaboration of Susan Barber, Elizabeth Rosenfeld, Zara Berro, and Matthew White Palmer.

References

Ardoin, S., Christ, T., Morena, L., Cormier, D., & Klingbeil, D. (2013). A systematic review and summarization of the recommendations and research surrounding CBM of oral reading fluency (CBM-R) decision rules. *Journal of School Psychology, 51*(1), 1-18.

Council of Chief State School Officers & National Governors Association. (2012). Supplemental information for Appendix A of the Common Core State Standards for English language arts and literacy: New research on text complexity. http://www.corestandards.org/assets/E0813_Appendix_A_New_Research_on_Text_Complexity.pdf

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

Cummings, K., Biancarosa, G., Schaper, A. & Reed, D. (2014) "Examiner error in curriculum-based measurement of oral reading". *Journal of School Psychology* 52, 361–375.

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

Deeney, T. (2010). One-minute fluency measures: Mixed messages in assessment and instruction. *The Reading Teacher*, 63(6), pp. 440-450.

Deeney, T., & Shim, M. (2016) Teachers' and Students' Views of Reading Fluency: Issues of consequential Validity in Adopting One-Minute Reading Fluency Assessments. *Assessment for Effective Intervention. Hammill Institute on Disabilities*, pp.1-18. DOI: 10.1177/1534508415619905

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

Eason, S., Sabatini, J., Goldberg, L., Bruce, K., & Cutting, L. (2013). Examining the relationship between word reading efficiency and oral reading rate in predicting comprehension among different types of readers. *Scientific Studies of Reading, 17,* 199-223.

Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology. 32: 221-233. Doi: 10.1037/h0057532.

Francis, D., Santi, K., Barr, C., Fletcher, J., Varisco, A., & Foorman, B. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46*, 315-342.

Goffreda, C. T., & DiPerna, J. C. (2010). An empirical review of psychometric evidence for the Dynamic Indicators of Basic Early Literacy Skills. *School Psychology Review*, *39*(3), 463.

Good, R. H., Kaminski, R. A., Cummings, K., Dufour-Martel, C., Peterson, K., Powell-Smith, K., & Wallin, J. (2011). *DIBELS next assessment manual.* Eugene, OR: Dynamic Measurement Group.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Accuracy and fluency in list and context reading of skilled and RD groups: Absolute performance levels and sensitivity to impairment. *Learning Disabilities Research and Practice, 18*, 237-245.

Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly, 45,* 230-251.

LaBerge, D., & Samuels, S.J. (1974). Toward a theory of automatic information process in reading. *Cognitive Psychology, 6,* 293-323.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common core state standards.* Washington, DC: National Governors Association Center for Best Practices, Council of Chief State School Officers.

National Institute of Child Health and Human Development, NIH, DHHS. (2000). Report of the National Reading Panel: Teaching Children to Read: Reports of the Subgroups (00-4754). Washington, DC: U.S. Government Printing Office.

Pearson, Inc. (2012). *Amisweb technical manual.* Bloomington, MN: Pearson. Retrieved at www.aimsweb.com/wp-content/uploads/aimsweb-technical-manual.pdf on September 28, 2017.

Rasinski, T.V. (2004). *Creating fluent readers.* Educational Leadership, 61(6), 46-51.

Schwanenflugel, P. J., & Benjamin, R. G. (2012). Reading expressiveness: The neglected aspect of reading fluency. In T. Rasinski, C. Blachowicz, & K. Lems (Eds.), *Fluency instruction, second edition: Research-based best practices* (pp. 35-54). New York, NY: Guilford.

Teacher's College Reading and Writing Project (2014). Teacher Resources and Guidebook for Levels A-K Reading Level Assessments. Retrieved from http://connect.readingandwritingproject.org/file/download?google_drive_document_id=0B7BccMltK6LqXy0waGdTV3QyaHM on 9/25/2017.

Teacher's College Reading and Writing Project (2014). Teacher Resources and Guidebook for Levels L-Z+ Reading Level Assessments. Retrieved from http://connect.readingandwritingproject.org/file/download?google_drive_document_id=0B7BccMltK6LqVi0talJ4TXNxZWs on 9/25/2017.

Wayman, M. M., Wallace, T., Wiley, H. I., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41*, 85-120.

Appendix A

The following are some statements. Read each statement. After each one decide how much you agree or disagree with the statement and mark your response.

		Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree
1.	Moby.Read was intuitive for kids.							
2.	The passages in Moby.Read were appropriate.							
3.	I was expecting something different for the scoring.							
4.	The voice was clear.							
5.	Instead of using Moby.Read, I would rather administer and score the test myself.							
6.	The visual layout of Moby.Read was pleasant.							
7.	I found Moby.Read frustrating to use.							
8.	The score information was useful.							
9.	I would prefer a different voice.							
10.	Running students through Moby.Read was more convenient than administering and scoring the test myself.							
11.	I will pass on using Moby.Read in the future.							
12.	The passages could have been better matched for my students.							
13.	Moby.Read could look better.							
14.	Overall, I liked Moby.Read.							