7th ISCA Workshop on Speech and Language Technology in Education
25-26 August 2017, Stockholm, Sweden

# Studies of a Self-Administered Oral Reading Assessment

*Jared Bernstein, Jian Cheng, Jennifer Balogh, Elizabeth Rosenfeld*

Analytic Measures Inc., Palo Alto, California, U.S.A.
Jared @ AnalyticMeasures.com, Jian.Cheng @ AnalyticMeasures.com

## Abstract

Reading assessments are most useful when they inform instruction. A prototype mobile-device app called Moby.Read presents a short self-administered test of oral reading fluency. Children read text passages aloud. Spoken responses are scored automatically on the device and displayed for teacher review. First, we report a usability study and a preliminary validation study with 99 children. Usability results are positive and score accuracy is high. Second, we discuss how accurate automatic reading assessment enables new analyses of reading performance (e.g. rate trends over time and within passage) that have not previously been available to guide individual instruction. We present early results of sub-passage and cross-passage results that hold diagnostic promise and discuss their potential to guide reading instruction.
**Index Terms**: oral reading fluency, comprehension, speech recognition, resilience, decoding, diagnosis, rate trend.

## 1. Introduction

Reading aloud is a convenient method for tracking progress in early reading. Informal reading inventories and oral reading fluency (ORF) tests yield observable evidence of a person's reading processes [1]. ORF procedures are used to benchmark children's reading in the early grades (ages 5-11), and ORF is often used to track progress in response to an instructional intervention. Although accurate automatic scoring of oral reading has been reported [2,3,4], teachers still spend about 4 million hours each year in the U.S. on administering and scoring basic reading assessments. Although ORF assessment is widespread, teachers and reading researchers have identified several problems including:

(a) inefficient use of teacher time [5];
(b) human scoring error from training deficits [5]
(c) emphasis on speed instead of expression of meaning [6];
(d) teacher uncertainty about how to use ORF scores [7]; and
(e) score instability from passage & individual factors [8],[9].

To address these problems, we developed an automated ORF assessment called Moby.Read, which is designed to improve the efficiency, specificity, and consistency of ORF measurement. Students take Moby.Read tests on a tablet computer, and it automatically scores oral readings and stores the recordings, which can be reviewed on the device.

Students self-administer the test and scores are reported automatically online. The app uses on-device speech recognition and scoring algorithms to return accurate reading rate (or words correct per minute: WCPM). Moby.Read's automatic scoring will reduce the need for teacher training and will help ensure consistency. Moby.Read's scoring of comprehension and expression will emphasize reading for meaning instead of reading for speed, and the reporting will help teachers link assessment results to instruction. Finally, Moby.Read's psychometric design will ensure better tracking of progress due to improved reading ability instead of spurious variability across passages.

## 2. Prototype App: Moby.Read

The prototype Moby.Read used in this research runs on an iPad as a stand-alone app. To get to the assessment, a simple, local ID is entered in the home screen. Entering a student ID causes Moby.Read to present one test session. When a Teacher ID is entered, Moby.Read will take the user to the Teacher Page where scores, trends, and recordings are reviewed.

The first time the student is taking a test, built-in video instructions are presented, explaining what to do and showing examples of a student performing the tasks. In each test ***session***, students are asked to read a word list, read an easy practice passage, and read three grade-level passages. After reading the practice passage aloud, students are asked to retell the passage and answer two questions aloud. Figure 1 shows screenshots of the Moby.Read function flow for a practice passage.

After the second comprehension question, students are asked if they would like to listen to a reading of the passage and try reading it again. If the student hits "yes," then the app presents the passage text again and plays a clear, expressive recorded reading of it. The text changes color in sync with the recorded reading, so children can easily follow along. Then the student is presented with the text again for a second reading (although this reading is not included in the overall score). Struggling readers liked this function because they gained confidence in their performance. Only the first, unpracticed reading is used in the overall score.

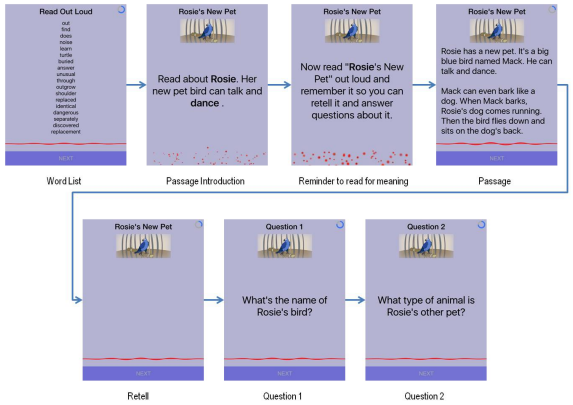


Figure 1: *Student flow for practice and test passages.*

After the practice passage, the student is presented with three grade-level passages (each followed by a request for a retelling, two comprehension questions, and the option to listen to the passage and read it again).


10.21437/SLaTE.2017-30

Figure 2 shows the teacher functions in Moby.Read. The teacher enters a Teacher ID to get to these pages on the device. The first page displays student names and IDs, and basic Oral Reading Fluency measures (rate, expression, comprehension). Selecting "View" activates the Graph screen, which shows a chart of the student's performance over time.

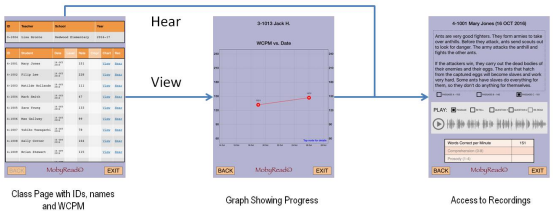


Figure 2: *Teacher pages: class table, progress graph, text/audio access.*

Selecting "Hear" on the class page activates the Recording screen where teachers can listen to all oral responses (readings, retells, and Q-answers), view texts, and see scores.

The Moby.Read design has had several iterations, based on expert and user feedback. First, existing reading fluency assessments were surveyed and reviewed for flow, content and reporting structure to see what teachers and administrators would expect. A storyboard served as a catalyst for discussion of the product and feedback from several reading experts and teachers. This feedback led to several changes in the design.

## 3. Preliminary Usability and Validation

To understand the usability and accuracy of the Moby.Read prototype, a pilot validation study was conducted and in-situ usability tests were run in several classrooms. Two studies addressed five research questions.

**A.** Can students in grades 2,3,4 (age 6-10) use Moby independently?
**B.** Do students have a good experience with the app?
**C.** Are rate & accuracy scores similar to scores by human listeners?
**D.** Are scores similar to human+paper administered ORF scores?
**E.** Do teachers find the app useful, intuitive and/or convenient.

Study 1 was a usability study in which facilitators provided iPads with the Moby.Read app to classrooms and students ran the app. Usability data were collected from both students and teachers. Study 2 reports on 20 of the 94 students, who each had an adult administer the DIBELS Oral Reading Fluency (DORF) section of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) *NEXT* assessment [10], a widely used human-administered oral reading fluency assessment. Assessment order was balanced across students.

Participants in Study 1 were 99 school-aged children from four different elementary schools: two in New Jersey and two in 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%). Of the 99, 51% of the students were European-American, 19% were African American, 4% were Asian, and 25% were Hispanic or Latino. Twenty of these students also participated in Study 2. Four teachers gave usability feedback via a teacher questionnaire.

For Study 1, the 99 experimental sessions occurred during the normal course of a school day at the participant's elementary school. Of the 99 student participants, 94 were able to run the App successfully. Several usability measures were gathered during the session: task completion, accurate reading rate in WCPM from the Moby.Read app, and a student judgment of how easy the app was to use (by selecting one of four *emoji* faces). After the sessions, teachers viewed score reports and could listen to student readings and retells. Then the teacher filled out the Teacher Questionnaire, which was a printed Likert-style survey.

For Study 2, a subset of 20 students took both a Moby.Read assessment and a DIBELS assessment. We used the fall benchmark form of DIBELS, which consists of three grade-leveled passages of about 250 words each. DIBELS administration and scoring followed procedures described in the test's official documentation [10]. We collected a Moby.Read WCPM score and a DIBELS Next WCPM score for each of the 20 students. Section 3.1 reports Study 2 results.

### 3.1. Usability and Validation Results

**A**: Can students in grades 2, 3, and 4 use the app independently? **Yes.** Of the 99 students who attempted an assessment with the prototype Moby.Read app, 94 were able to go through the app and provide responses that were scored and were usefully made available to the teacher. Of the five students who were not able to provide data, two had technical problems with the ear phones (not plugged in) and did not hear the audio instructions; and three spoke too softly or read silently.

**B**: Do students have a good experience with the app? **Yes.** The usability scale ranged over 1-4, as: *totally confused* = 1, *not sure* = 2, *I knew what to do most of the time* = 3, and *easy* = 4. Responses ranged from 2 to 4 (mean: 3.4), indicating most of the 94 students found Moby.Read easy to use and almost all knew what to do as the assessment progressed through the tasks. In Study 2, each of the 20 students was presented with a forced choice question: *Which one did you like better: the version on the iPad or on paper?* Of the 20, 18 said *iPad* and 2 said *both*.

**C**. Are Moby's rate & accuracy scores similar to scores judged by human listeners? **Yes**. Each response from the 94 student sessions was human scored by two expert raters with high inter-rater correlation. 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.

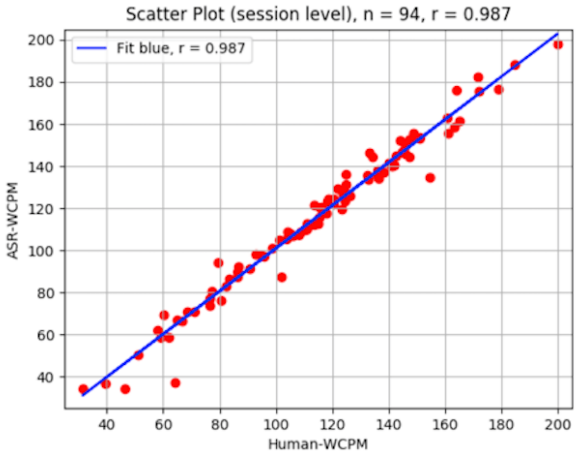


Figure 3: *Session-level scatter of median WCPM; server-based scores vs. Human scores (r=0.987).*

The human WCPM in Figure 3 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.

**D**. Are Moby.Read scores similar to human+paper ORF scores? **Yes.** (r=0.88) The correlation found in Study 2 between Moby.Read scores and DIBELS NEXT scores was 0.88. Published studies of DIBELS report a test-retest reliability of 0.82 and an inter-rater reliability of 0.85 [11]. The reliability of an instrument limits the strength of the correlation between that instrument and other measurements. So, the correlation with Moby.Read is at the ceiling of what would be expected, given the reliability of DIBELS.

**E**. Do teachers find the app useful, intuitive and/or convenient? **Yes.** The teacher questionnaire assessed opinions as to how useful, intuitive, or convenient Moby.Read was, and responses were quantified on a seven-point scale. Responses: The average rating was 6.1, which suggests that teachers found the app useful, intuitive and/or convenient.

### 3.2. Acoustic Models

The acoustic model AMI used for speech recognition on device is a DNN-HMM one with four hidden layers, trained using Librispeech [12] 961 hours of clean native (L1) reading data. Moby.Read's sample rate is 8,000. The DNN acoustic features are from a 40 Mel-scale filter-bank (FBANK). We concatenated a context of six feature frames left and right of the current frame (13 frames total). The number of senones (tied states) is 2064.

There are several channel mismatch issues that may degrade the ASR performance: 1) an adult acoustic model was used to recognize children data; 2) we used narrowband; 3) the acoustic model used here was trained using very clean/quiet recordings, so Moby.Read scoring accuracy may diminish with very noisy data. Although we note these potential issues, the overall on-device acoustic model performance seems good.

## 4. New Analyses of Oral Reading

Automated scoring based on speech recognition opens new windows on reading processes. Sub-passage measurement enables us to see student reading performance in word-by-word detail. Reading rate can be calculated automatically for each word, or phrase, or window of N words in a passage. If we can automatically track reading rate fluctuations over the words and phrases in a passage, we can automatically diagnose which structures are most challenging for a student. If we collect and analyze a suitable sample of readings of a given passage, we can normalize that diagnosis with reference to passage and peer group.

The usual procedure in measuring oral reading fluency is to present three passages to read aloud and then report the median WCPM value from the three passages. Some readers are more familiar and fluent readers on one topic than on another. There are significant interactions in reading performance between individuals and passages, so taking a median WCPM value makes sense. However, this practice ignores the possibility that there may be an overall trend in WCPM within and between passages, and that there may be other valuable diagnostic information within each passage. For example, a reader may give a fluent reading of a passage with prosody that reflects and expresses meaning, but gets stuck on one word.

Especially for "struggling" readers who are reading below grade level, their accurate reading rate may decrease as a passage continues, or as they go from one passage to the next. Such a decrement may be caused by difficulty with particular words or topics, or by reading-fatigue that may cause some struggling readers to "give up" on the task and continue with a lowered level of active attention.

We report here on just one diagnostic measure of reading performance: Resilience (or perseverance). In particular, we seek evidence to support the hypothesis that some struggling readers may show a negative trend in accurate reading rate as Moby.Read's passages continue or as the passage-after-passage administration continues. Such a negative trend is observed directly in reading rate (WCPM), but will be best understood when normalized against the rate trends of a cohort of readers.

## 5. Rate Data & Analysis

Each of 94 students with reportable data (24 in grade 2, 38 in grade 3, 32 in Grade 4) read three passages aloud in a Moby.Read assessment. For each passage reading from each student, we aligned the words in the oral reading response transcript with the original passage text by using a standard minimum edit distance. Then we moved a 10-word window over the passage source text, progressing through the text in 5-word shifts. So, for example, in the *Zack-Backpack* passage analyzed in Figures 4 and 5, there are 16 text spans. At each text span, when eight or more of the 10 source-text words were read aloud by that student in a correct order, we calculated that window's read-aloud duration to compute WCPM in that window. If fewer than eight words (of the 10 in the source-text span) were read aloud, then that window was skipped and had no WCPM value. The window shifted to the right by five words and computation was repeated until we reached the end of the passage, producing a sequence of 16 WCPM values centered at every fifth word in the 89-word *Zack-Backpack* passage.

This running WCPM sequence has potential value for the struggling readers, as it yields information that locates the points in the text where readers are having trouble relative to their peers. This is of use in editing and preparing texts, as well as in providing formative feedback that teachers can use.

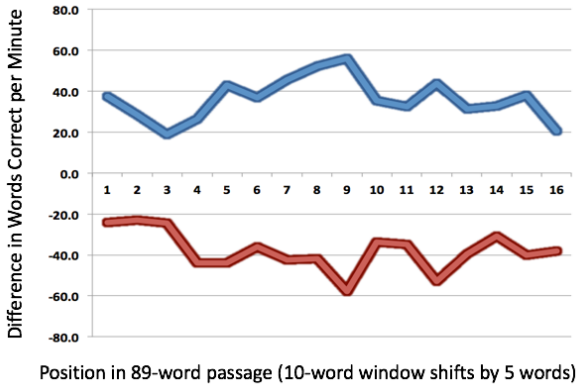


*Figure 4: Sequences of difference values of accurate reading rate (WCPM) for 10-word text-spans in one third grade story. At each point on the X axis, zero on the Y axis is the WCPM value averaged over all 38 3rd graders at that place in the passage.*

In Figures 4 and 5, the indices 1 to 16 are text spans. Span 1 is centered on word 5 and includes words 1-10; span 8 is centered on word 40 and includes words 36-45; span 16 is centered on word 80 and includes words 76-85. The upper line in Figure 4 shows average WCPM for the 11 fastest readers minus the

average value for that 10-word span across all 38 of the 3rd grade students. The lower line is the average value for the 11 slowest readers minus the overall averages. The source text for Figures 4 and 5 is one of 9 texts used in this study. It reads:

*Zack sat down in his chair. It was time for class. The teacher asked for last night's homework. Zack opened his backpack to take out his homework folder. His homework folder wasn't in it. His recorder was gone too.*
*In the backpack there was a science book and a math book. They were both too hard for Zack. There was a piece of homework inside a green folder. The name on it said 'Jason'. Zack must have taken his brother's identical backpack while hurrying out the door this morning.*

Data displayed in Figure 4 suggests that for relatively slow readers, the two hardest spans of text in this passage are #9 near word 45: *…the backpack there was a science book and a math…* and #12 around word 60: *…hard for Zack. There was a piece of homework inside…*

## 6. Results

There are many hypotheses that can be addressed with the within-passage data from Moby.Read, but for now we focus on questions about differences between faster and slower early readers. In particular:
Do slow readers show evidence of fatigue as a test progresses?
Is rate variation different for slow and fast readers?

This data indicates that early readers may slow down over three passages that total about 270 words. Across all grades and readers, the average reading rate is 2% faster in the second passage, then 5% slower in the third passage.

Do slow readers slow down during a passage? Do they show differential slowing compared to faster readers? Analyzing data only from the 38 third graders who read *Zach's Backpack*, data suggest that all readers generally slow down a bit over the course of a passage, and that the slower readers slow down much more. Using best-fit linear slopes, the 11 fastest readers (Fig. 4 upper line) lose about 4% of their reading rate over the length of that passage. On average, the 11 slowest readers lose about 54% of their reading rate over that same passage. The WCPM trends are very different, although Figure 4's averaging and aspect ratio obscures this difference.

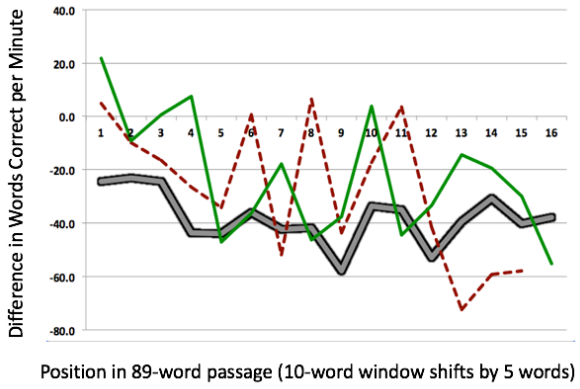


*Figure 5: Sequences of difference values of WCPM for the text-spans. Zero is WCPM value averaged over 38 3rd graders at each span in the passage. The heavy gray line is the average WCPM for the 11 slowest readers. The thin solid and dashed lines show rates from two individuals among those 11 slow readers.*

To examine the variation in reading rate, we calculated the standard deviation of the reading rates for each of the fastest and slowest readers, expressed as a percentage of that reader's average rate. For the slow readers, the standard deviation averaged about 32% of their rate; for the fast readers, it averaged about 21% of their rate. An example of differential reading rate for two slow readers against the average reading rate of the slow group is shown in Figure 5. Note first that the rates of these two slow readers varies considerably from span to span, even though adjacent spans overlap. Second, note that the points of relative difficulty in the text are quite distinct for these two readers, which suggests that their sub-skill profiles are also likely to differ.

## 7. Discussion of New Analyses

This partial analysis of reading rate trends is consistent with the hypothesis that many struggling readers get discouraged and perform below their potential later in a given passage and in the later passages among a set. The finding of greater variability in rate for slower readers was expected and may offer the main key to providing useful formative information to teachers.

Note first that the passages were specified and written to increase in difficulty from beginning to end, so some part of the within-passage rate decrement is probably due to the passage structure. Second, note that the readings were presented in a fixed order to the 94 students whose data was analyzed, so the passage-to-passage rate decrement may partly be an artifact of differences in items difficulty.

## 8. Conclusions

Evidence from this small sample suggests that about 95% of U.S. school children (as young as 6 or 7) can successfully self-administer an oral reading fluency assessment as implemented in Moby.Read. Automatic scoring of these self-administered assessments correlated very highly with double human WCPM scorings of the same reading performances. The automatically scored Moby.Read tests predict and align with concurrent scores from standard, commonly used, adult administered ORF tests.

Preliminary results from sub-passage and inter-passage analyses suggests more (and different) instruction-relevant data can be automatically extracted from oral readings than has traditionally been inferred from measures of accurate reading rate (WCPM).

## 9. Acknowledgements

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.

## 10. References

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

[2] Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., Huang, C., Junker, B., Sklar, M. B.,& Tobin, B. (2003). Evaluation of an automated Reading Tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research*, 29(1), 61-117.

[3] R. Downey, D. Rubin, J. Cheng & J. Bernstein (2011) "Performance of automated scoring for children's oral reading". Paper presented at the BEA workshop, NAACL-HLT 2011, June, Portland, Oregon.

[4] Balogh, J., Bernstein, J., Cheng, J.,Van Moere, A., Townshend, B., & Suzuki, M. (2012) "Validation of Automated Scoring of Oral Reading". *Educational & Psychological Measurement*, 72(3), pp. 435-452.

[5] 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.

[6] 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.

[7] 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

[8] 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.

[9] 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.

[10] Good, R. H., & Kaminski, R. A. (2011). DIBELS Next Assessment Manual. Eugene, OR: Dynamic Measurement Group (Retrieved from: http://dibels.org.).

[11] Goffreda, C. & DiPerna, J. (2010). An empirical review of psychometric evidence for the dynamic indicators of basic early literacy skills. *School Psychology Review*, 30(3), 463-483.

[12] V. Panayotov, G. Chen, D. Povey, S. Khudanpur (2015) Librispeech: An {ASR} corpus based on public domain audio books, in *Proc. IEEE ICASSP 2015*, South Brisbane, Australia, April 2015, 5206--5210