CHAPTER 3

ARTIFICIAL INTELLIGENCE FOR SCORING ORAL READING FLUENCY

Jared Bernstein
Stanford University

Jian Cheng
Analytic Measures Inc.

Jennifer Balogh
Intelliphonics

Ryan Downey
Analytic Measures Inc.

ABSTRACT

We describe assessments in which machine learning has been applied to develop automatic scoring services for constructed responses; specifically to score students' spontaneous spoken responses. We review a few current large-scale examples and then describe recent work with automatic scoring of an oral reading fluency (ORF) instrument that runs on mobile devices. We re-

Application of Artificial Intelligence to Assessment, pages 51–76

port data that verified the accuracy of the ORF scores and summarize student engagement and teacher responses to the ORF instrument.

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Artificial intelligence (AI) has promised to improve decision-making and to relieve the burden of tedious human tasks. In education, teachers face a tradeoff between offering more instruction time and devoting more time to evaluating student skills (to inform the instruction). In the United States, where measured outcomes (e.g., "adequate yearly progress" for the *Every Student Succeeds Act* of 2015) may determine school funding and teacher pay, the need to quantify students' performance has resulted in an environment where some teachers feel they spend too much time testing (cf. Common Core, 2011), leaving less time for instruction.

First, we differentiate *machine learning* (ML) from AI. Then, we consider how machine scoring relates to educational testing. Note that both ML and AI are relatively recent terms, so their meanings may change over the next decades. Briefly put:

AI refers to an automated activity that recently needed biological systems to be accomplished;
ML is the engineering field that develops technology to build AI systems.

ML refers to algorithmic processes that operate on data sets to produce algorithms that cluster, classify, or identify patterns in new (unseen) data sets. Typically, an algorithm infers a function (or a model) that assigns labels to new data from an analysis of labeled training data. For example, thousands of transcribed voice recordings are analyzed by an ML procedure, which produces an algorithm that transcribes new voice recordings.

Linear regression from one independent variable to predict a dependent variable is an ML method that has been in use for 200 years and that will be familiar to many readers. Newer ML algorithms implement many more complex statistical models and methods, including more familiar techniques such as logistic regression and decision trees, and others such as Bayesian models including hidden Markov models (HMM) which are commonly used in sequential pattern recognition. New statistical models are appearing continuously; recent ones include support vector machines (SVMs) and deep neural networks (DNN) used for classification, and many other models and methods that are emerging every year. These techniques can be combined and applied to data to support classification or predicting continuous variables (e.g., expected time to task completion).

Artificial Intelligence is an ability of automated systems to perform tasks that *until recently* required human or other biological information processing,

often including sensory perception, decision, and sometimes, verbal and/or mechanical response. With the wide adoption of any new ML technologies, the AI label seems to fade. What was AI 10 years ago, is now accepted without question. Currently, in 2020, systems that effectively search within vast bodies of text are commonplace, but 10 years ago, effective search required "AI" to produce accurate results. This year, autonomous vehicles seem magical, and they are now seen as AI systems, but 10 years from now, they may seem to be just computer-based systems that rely on well-understood computational methods. The usage of these terms is fluid and may crystalize in unforeseen ways. Machine learning has already had an impact on assessment and there are more applications underway.

ASSESSMENT

Assessment is the evaluation of a test taker's ability, skills, or knowledge. It has several elements. The process of assessment (see Figure 3.1) involves identifying the test taker (security, proctoring) and presenting content (items, test forms) to the test taker (administrative platform). This platform also collects test taker responses (e.g., selections, completions, written or spoken performances, or other behavior samples), and the responses are scored and reported. Finally, the results of an assessment are consumed by the score user, often a teacher or administrator, to satisfy a decision demand (e.g., refine curriculum, apply an instructional intervention, pass/fail, rank for acceptance).

AI Impact on Assessment

The *purpose* of an assessment depends on the decisions that are made based on test results. In education, assessments often quantify some aspect of a learner's state, with the goal of enabling test score users to make

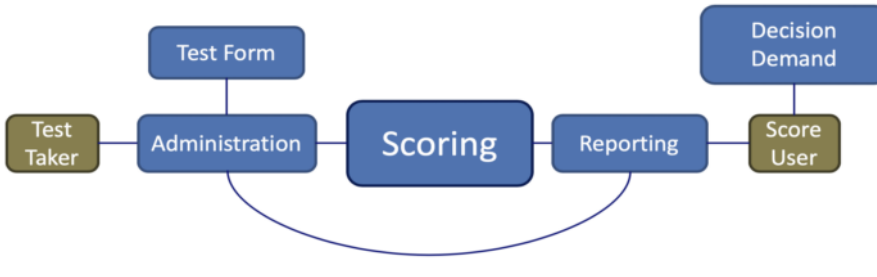

Figure 3.1 Select components in assessment.

warranted decisions about the learner or about the efficacy of an instructional approach. Examples of educational decisions include:

- Has the test-taker mastered the content of a particular lesson? (curriculum mastery)
- Does the test-taker exhibit reading skills adequate for success? (placement, screening)
- Has the test-taker reached skill levels needed to succeed in college? (admission, achievement)
- Do the test-takers meet curriculum goals? (program evaluation, accountability)

To ensure that a test score reflects an evaluation of the performance that is needed to inform a particular decision, test tasks should elicit behavior that resembles the test-taker's actions in the target situation. For example, a test that intends to evaluate a candidate's conversational skills should include tasks that themselves resemble conversations, or should require the candidate to produce spontaneous speech. With the advent of optical mark recognition technology in the 1930s, machines were able to score items as correct or incorrect, which dramatically increased scoring efficiency and testing volume. However, this technology also shifted the focus of testing to receptive skills like listening and reading, or to more discrete subskills such as vocabulary, because those skills are easier to test by selection and score by optical mark recognition.

American K–12 students are increasingly asked to demonstrate proficiency in academic skills using constructed-response tasks. As students mature, they are expected to move beyond merely demonstrating knowledge of academic facts and are expected to be able to apply knowledge in productive ways. This trend is driven by the spirit of the Common Core State Standards (CCSS), which acknowledges that, to be successful in college or in a career, students must synthesize and produce complex material, rather than just select the correct answer from among a given set (CCSO & NGA, 2012).

Some new assessments that align with CCSS use "technology-enhanced items" (TEIs), with response types that include "drag and drop," "multiple select," or "hot spots." These digitally enabled response types can be somewhat more constructive, informative, or authentic than traditional selection-response item types. However, such items usually still produce dichotomous scores and do not enable students to demonstrate productive control of integrative, complex skills such as an ability to "propel conversations by posing and responding to questions that relate the current discussion to broader themes or larger ideas; actively incorporate others into the discussion; and clarify, verify, or challenge ideas and conclusions" (CCSS. ELA-Literacy.SL.9-10.1.c).

Practicality also informs test design. Because constructed-response items have traditionally required human scoring, the operational burden of human scoring (often double scoring) has resulted in a testing culture in which large scale assessments design the vast majority of test tasks for multiple-choice responding.

Given the need to demonstrate academic progress for some 55 million K–12 students in the United States, some states have invested in automated scoring technology to evaluate productive skills (e.g., speaking) in tests such as the Arizona English Language Learner Assessment (AZELLA; see Cheng, Zhao, D'Antilio, Chen & Bernstein, 2014) or the Texas English Language Proficiency Assessment System (TELPAS). Figure 3.2 schematizes a typical tradeoff in costs associated with human scoring versus machine scoring constructed response items and performance-based items in a hypothetical test.

As shown in Figure 3.2, automated scoring involves higher up-front costs. After one develops test tasks that elicit constructed responses that can be scored reliably by human raters, one still has to run training procedures that generate accurate algorithmic scoring models for the constructed-response items. However, with time, the cumulative cost of the automatically scored testing should be much less.

Automatic scoring is also more practical because scores generally come back immediately. In a formative, classroom-based context, a teacher needs to know what areas of strength and weakness a student has, or whether the instructional material has been effective; the sooner the teacher has this information, the sooner appropriate intervention can be introduced. (Traditional formative classroom assessment was a simple poll, such as: "Raise your hand if you think it's answer A.")

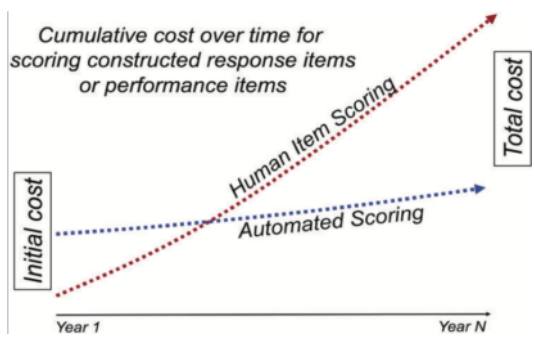

Figure 3.2 Cumulative cost over time; higher initial cost for automated scoring combined with lower recurring costs; lower development cost for human scored tests with higher recurring costs.

Rapid turnaround of scores for higher stakes assessment means the candidate can more quickly decide whether to retake the test. For example, candidates taking certain high-stakes tests may experience a delay of days or weeks before their scores are released, largely due to the need for human scoring of the written and spoken portions. Recent tests, such as the Pearson Test of English-Academic, can guarantee that scores are returned within 3–5 days because speaking and writing are machine scored.

If ML procedures are applied to well-designed training data, computer scoring can reduce bias and help ensure fairness and reliability. Test tasks may need to be reworked to increase human rater agreement, and a larger, more diverse sample of candidates may be required. Most important, to establish reliable training targets, it often helps to train with scores from three or four human raters, rather than just one or two raters. To the extent that a digital system is trained properly, automatic scoring should produce accurate and consistent scores over time, and across location and candidate characteristics.

Thus, an automated scoring system that evaluates performances in constructed-response tasks may serve secondarily to expand the test developer's toolkit, increasing the range of performance types that can be scored affordably, immediately, and consistently. Automatic scoring systems need to produce scores that are consistent with, or that improve on, the scoring expected from trained human judges. Beyond simply replicating the scores that a human would produce, AI can expand the range of information used in an assessment.

Artificial intelligence has also changed educational practice and assessment demands. Grammar checkers and spell checkers use ML and natural language processing (NLP) techniques to identify ungrammatical constructions and flag unconventional word choices; some make stylistic recommendations that better capture what the writer (is predicted to have) intended. As spelling- and grammar-checking improves toward very high accuracy, these correction functions are being included in some testing platforms, as it may seem less important to assess accurate spelling and grammar conventions, because prescribed orthography and grammar are normally tracked and corrected in text processing interfaces. By analogy, when starting a gasoline engine, we no longer need to know how to retard the spark and choke the carburetor, and so these skills are not part of current driving exams.

Artificial Intelligence in Assessment of Language

Advances in ML permit the reliable scoring of new kinds of tasks, including performance-based tasks, which previously could only be scored by trained expert human judges. In recent decades, natural language

processing (NLP) techniques have made it possible for computers to generate scores for some aspects of written essays such that the scores agree well with scores from trained human graders (cf. Foltz, this volume). Note, however, that ML methods have had more success scoring more global qualities of student essays, for example, general content coverage and coherence, while some superficial, and seemingly easier, tasks like correcting grammar and usage are not yet accurate enough to be very helpful to either student writers or teachers. Automated scoring of writing has been applied in several commercial applications such as TurnItIn.com's plagiarism checker and Pearson's *WriteToLearn*. Van Moere and Downey (2016) provide an approachable description of the automated scoring technology underlying such applications.

Machine learning supports the automated assessment of speaking skills, as well. For example, automated speech recognition (ASR) converts an acoustic signal into a stream of feature vectors (about 100 vectors per second) that are then decoded to find the most likely word sequence that the ASR system is trained to hear. Limited applications of ASR have been included in language teaching systems since their introduction in 1995, by Syracuse Language Systems. Since then, Rosetta Stone, Duolingo, Babbel, and others have applied ASR in ways that are sometimes useful to learners. Several products also present visual representations of a learner's speech for comparison to visual displays of model native speech. Bernstein and Franco (1996) and Young (1996) provide relatively accessible descriptions of how speech recognition systems work. Both papers are a bit out of date, but at a first level of understanding, they are still accurate.

Automatic spoken response scoring may focus on the content of the speech, which includes its turn structure, pragmatic force, linguistic form, and lexical content. Scoring might also focus on qualities of the speech itself such as its fluency and pronunciation, or it may combine content and quality aspects into a more general estimate of speaking ability. Figure 3.3 shows the elements in a system that elicits and scores spoken responses from a student.

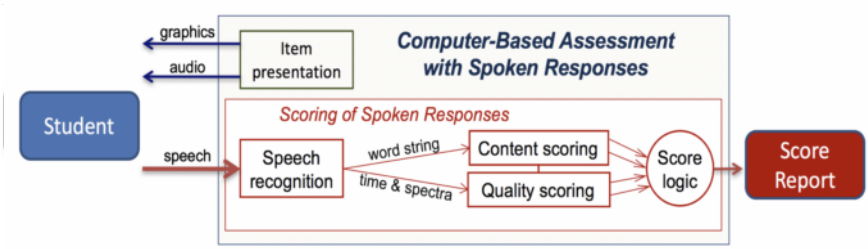

Figure 3.3 A system for presenting and scoring spoken-response test items.

Current computer scoring systems estimate speaking ability by combining measures of linguistic and lexical structures with measures of fluency and pronunciation, returning scores with high consistency and accuracy. How is it done?

When a digital recording of speech is made, the acoustic signal is transduced into an electronic signal and then converted into a digital signal. The samples in this digital wave file are grouped into small sets of adjacent samples, called *frames*, which extend over 10 or 25 milliseconds of signal. ASR technology extracts timing and frequency information from these samples and applies acoustic models, language models, and a dictionary (with expected pronunciations) to decode the speech based on the digital information. When the test-taker's words are recognized and aligned in time with the signal, the words-in-time are evaluated with reference to statistical models that represent various levels of performance to estimate which level a given performance belongs to. Bernstein (2012) describes the application of these technologies for spoken language tutoring and testing, and attempts to explain their limits. A fuller discussion of validity in automated assessment of spoken language can be found in Bernstein, Van Moere, and Cheng (2010).

Training and Validation

To have confidence in automatic scores, one has to ensure that machine learning produces accurate and consistent results. Large training data sets are required to ensure adequate representation of the variability expected in the target performances when the models are optimized. In addition, validation testing ensures that the scoring models appropriately evaluate unseen data. When using ML to support the assessment of human language, one usually needs to demonstrate that machine-derived scores are comparable to those assigned by trained human judges. This step is accomplished using a new sample of responses, previously unseen by the system during optimization. When available data is limited, an efficient way to predict the machine-human agreement for future (unseen) data sets is by partitioning all available data into 5 or 10 subsets and performing a cross-validation (see Hastie, Tibshirani, & Friedman, 2009; Chapter 7). However, if data is plentiful, then, at a high level, the steps required to train and validate an automated scoring system are schematized in Figure 3.4.

EXAMPLE IN DETAIL: ASSESSMENT OF ORAL READING FLUENCY (ORF)

To illustrate how ML is applied to build AI into an assessment, we describe the development and evaluation of an automated reading test. It performs

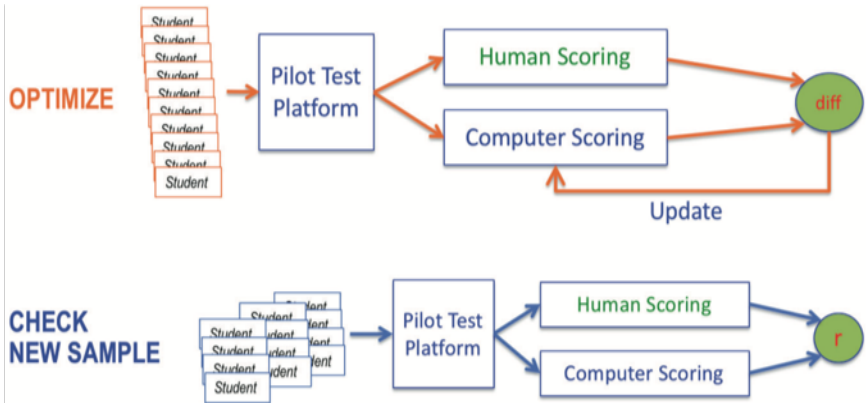

Figure 3.4 Optimizing the scoring system and confirming its accuracy on a new sample.

tasks that until recently required the attention and actions of a trained reading teacher. The first part of this chapter described how ML can enable automated scoring of spoken responses to complex linguistic tasks. In this section, we describe the development of one ORF test, Moby.Read®, that evaluates reading skill in young students.

Oral Reading Fluency

The National Reading Panel (2000) defined oral reading *fluency* as a reader's ability to "read a text quickly, accurately, and with proper expression." *Accurate rate* is the number of words read correctly per minute (WCPM). Note that WCPM can be obtained by collecting reading performances with durations shorter or longer than one minute. For example, a student correctly reading 55 words in 30 seconds has a reading rate of 110 words correct per minute. *Accuracy* of reading is an important measure because it reflects a student's skills in recognizing common words and in decoding letter sequences in unfamiliar words. A student who reads quickly but makes errors will receive a different accuracy score from a student who reads more slowly but with high accuracy, and these two patterns carry different information about the reader's comprehension of text and application of text-embedded information in real-life situations. Finally, *expression* takes into account appropriate pacing, pausing, syllable- and word-level stress and pitch, and other aspects of speech which differentiate a monotonous, non-expressive reading from a reading that embodies text meaning and/or structure (Schwanenflugel & Benjamin, 2012; Schwanenflugel, Hamilton, Kuhn, Weisenbaker, & Stahl, 2004).

All three components of ORF (rate, accuracy, and expression) have been shown to correlate with reading comprehension (Daane, Campbell, Grigg, Goodman, & Oranje, 2005; Fuchs, Fuchs, Hosp, & Jenkins, 2001). Students who read aloud with high accuracy also tend to score high on measures of reading comprehension.

Oral reading fluency has been used for decades as a reliable reflection of early reading ability (e.g., Shinn, Good, Knutson, Tilly, & Collins, 1992), as well as an indicator of a student's general academic level (Hasbrouck & Tindal, 2006), particularly in Grades K–6, when core reading skills develop. Available standardized ORF measures are popular because they are relatively reliable and brief. These include *aimswebPlus* (Pearson, 2018a), *Dynamic Indicators of Basic Early Literacy Skills* (Dynamic Measurement Group, 2018), and *EasyCBM* (Houghton Mifflin Harcourt, 2018), among others. Typically, students are asked to read three separate passages, reading as much of the passage as they can in 1 minute, to establish a reliable baseline or benchmark. For monitoring progress during an instructional intervention program, a single 1-minute reading is often accepted as sufficient. A teacher times the reading while following along with the student, annotating another copy of the passage for errors (e.g., word omissions, substitutions, transpositions). Once the readings are finished, the teacher tallies up the errors and records accuracy and reading rate. To enhance score stability, the median score across the three passages may be selected as the final record for each trait. A judgment of reading expression is sometimes a part of the official record.

A complete profile of the reader also includes a measure of comprehension (e.g., Deeney, 2010). For example, the CCSS (CCSSO, 2010) include reading comprehension in its foundational skills in the English language arts standards, which state that students should be able to:

- "Read with sufficient *accuracy* and fluency to support *comprehension*" (CCSS.ELA-LITERACY.RF.2.4).
- "Read grade-level text orally with *accuracy*, appropriate *rate*, and *expression* on successive readings" (CCSS.ELA-LITERACY.RF.X.4.B).

Despite their widespread use, there are drawbacks to traditional ORF measures. To score ORF performances, administrators have to learn annotation conventions to properly categorize reading errors and learn what to do when unusual performances are observed (e.g., when a student skips an entire line of the text, or when a student gets stuck on an unfamiliar word). Because a teacher follows along and annotates the reading in real time, oral reading performances are administered one-on-one, taking the teacher away from the rest of the class for 20 to 30 minutes per assessment. Teachers often skip rating the read-alouds for expression because reporting

expression scores is not required, and many teachers are unsure of their skill in making such a judgment when they have concentrated on reporting rate and accuracy (Schwanenflugel & Benjamin, 2012).

Development of Moby.Read

The Moby.Read assessment was developed to provide teachers with an easy way to get accurate ORF measures for children in Grades 1 through 5. The Moby.Read assessment measures four components of reading skill (comprehension, accuracy, reading rate, and expression) on leveled text for children in Grades 1–5. In addition, the Moby.Read app also reports an overall Moby.Read level that integrates these separate components into a reader-level estimate. The Moby.Read test was developed to be taken on touchscreen-enabled devices to facilitate ease of interaction for the youngest students and to enable on-device speech recognition technology to provide scores and feedback immediately, that is, without the need for a teacher to follow along annotating and tallying errors.

Test Structure

The Moby.Read test has several sections (see Figure 3.5). First, students watch an *instruction video* explaining the test and showing a student taking the test and providing spoken responses. The video introduces the test sections and provides a model for students, reminding them to read for meaning. At the end of the video, the student reads a *word list* out loud, and then

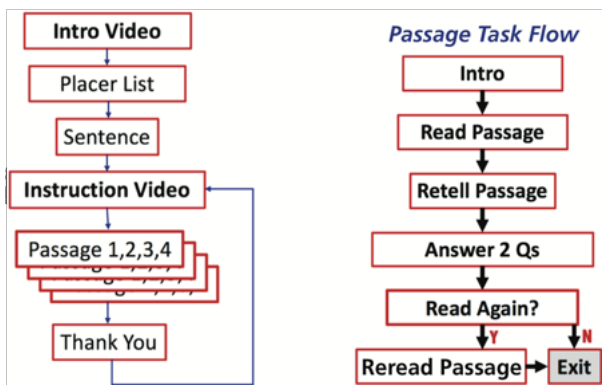

Figure 3.5 Left: Overall task flow during a Moby.Read session. Right: Per-Passage task flow.

reads a *sentence* out loud. These responses can be used to help estimate the student's reading level.

The passage-reading section starts with a video that shows an example of a student reading a passage out loud, retelling the passage, and then answering comprehension questions. Beginning with an initial unscored practice passage, the test flow consists of three test blocks, which proceed as follows: Students read a passage on the screen out loud; then students retell the passage aloud in their own words from memory; finally, students answer two comprehension questions about the passage (see Figure 3.6). In all cases, the response is spoken out loud by the student. Students who wish to read the passage again may select that option, hear a fluent reading of the passage with synchronized text highlighting. If they do not wish to reread a passage, they tap the screen to advance to the next passage.

Moby.Read passages are written to conform with the two broad text types described in the CCSS: (a) literature such as stories and (b) informational texts that cover topics such as history, art, social studies, science, and technology (CCSSO & NGA, 2012). For appropriate domain sampling, each form of the Moby.Read test includes both narrative stories and informational texts that provide facts and background knowledge on selected topics. Passages are leveled using the method prescribed by the CCSS (as described in the test development section). Passages contain 40–145 words,

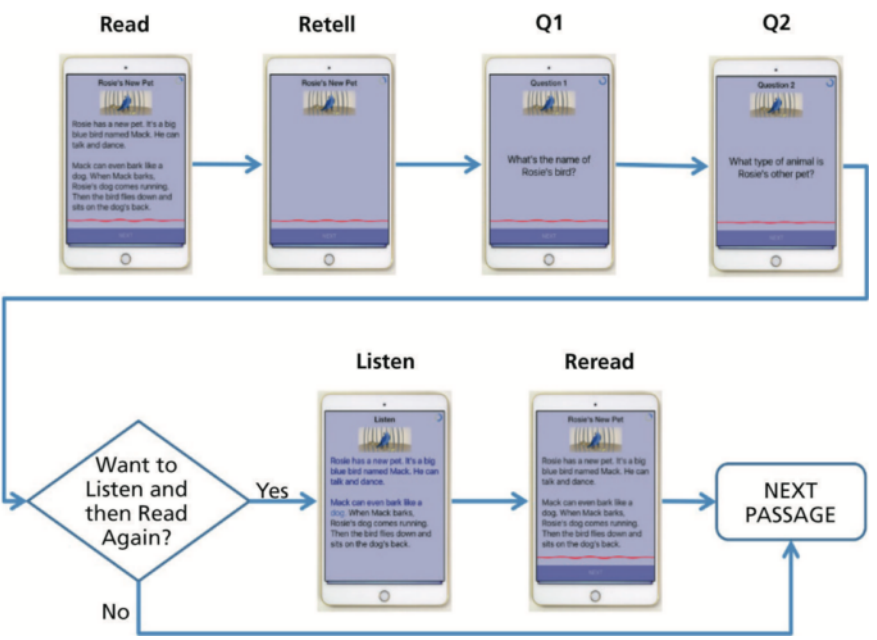

Figure 3.6 Test block flow for Moby.Read.

designed to provide enough spoken material for reliable scoring while being short enough to allow for efficient testing and completion in a reasonable time by the typical student given the grade level (Hasbrouck & Tindal, 2006). For empirical leveling, 376 students participated in studies piloting a total of 82 passages. Based on student performance in these studies, passages that were too hard or too easy compared to the passage's assigned Common Core level were assigned either one grade higher or lower, or were removed from the item pool.

The decision to include a passage *retelling* task is twofold. First, retelling a story requires that the reader understood the content of the passage. The extent to which the retelling captures key elements of the reading is a reflection of the reader's comprehension. Further, retelling is an established strategy for encouraging deeper processing of text (Morrow, 1985; Wilson, Gambrel, & Pfeiffer, 1985).

Comprehension questions elicit responses that might be a single word, a short phrase, or a sentence. Passage questions range in their cognitive processing requirements, from literal questions about facts in the text to questions that require the intersection of facts and/or inference. Most Moby. Read comprehension questions are direct and literal.

Speech Recognition System

Before applying the algorithms that produce specific oral reading scores, the child's recorded reading is analyzed by Moby.Read's ASR system. The acoustic model used for the ASR system was a Deep Neural Network-Hidden Markov Model (DNN-HMM; Zhang, Trmal, Povey, & Khudanpur, 2014) with four hidden layers. Language models for reading responses were rule-based and are specific to each reading passage. Language models for retelling responses and comprehension questions are also item-specific. Given that the language models are item specific and that correct readings are generally more likely than incorrect readings, the ASR system gives the reader the benefit of the doubt and will usually accept that a child has said a word correctly, even if the child's speech is accented. The Moby.Read ASR engine is based on Kaldi (Povey et al., 2011), and has been optimized for children's speech and to run in an iOS device. Jurafsky and Martin's book (2009) is a reliable textbook introduction to speech and language processing that covers most of the underlying ML methods discussed below. At this writing, a .pdf version of the forthcoming 3rd edition is available at https://web.stanford.edu/~jurafsky/slp3/ and can be downloaded from there.

Time-aligned response strings generated from the ASR system are scored with respect to the read text. For example, timing information at syllable,

word, and phrase level, along with inter-word silences, are used in expression scoring.

Models for Automatic Scoring

Automated scoring models were trained on reading performances by 383 students in Grades 2 through 6 in California ($n = 261$) and New Jersey ($n = 122$). The data-collection tests were automatically administered on an iPad with children wearing headsets with inline microphones. The tests advanced as appropriate based on the system's detection of the student's speech or screen touches. Typically, a human proctor monitored several students in concurrent sessions, with students sitting more than 1 meter apart.

Accuracy

Accuracy quantifies the percentage of words the reader decoded and spoke correctly. For accuracy scoring, the response string produced by the ASR system is aligned to the passage text to determine the first word attempted by the reader and the last word attempted. The alignment in combination with the language model allow an algorithm to be developed that calculates the number of words the student read correctly. Self-corrections are correct, but omissions, substitutions, and radical mispronunciations are not counted as correct. The accuracy score is reported as a percentage of words read correctly over the number of words attempted.

Accurate Reading Rate

For each scored passage, the words read correctly between the first word attempted and last word attempted are tallied. This tally is then divided by the time between the onset of the first word attempted and the offset of the last word attempted. Accurate reading rate uses WCPM as its unit, with the median rate over three passage texts reported to users.

Comprehension

Comprehension is the degree to which a reader can identify or present the major and minor concepts, themes, and/or facts contained in a passage. Moby.Read comprehension scores are reported to score users on a 0.0–8.0 scale, with higher values representing greater comprehension. Comprehension scores are derived from scores of the *retellings* and scores on *comprehension questions* combined across passages. The reported comprehension score gives equal weight to the retellings and to the answers to comprehension questions, averaged across the three passages.

The scoring algorithm for *retelling* measures the semantic similarity between the passage text and the retelling response using trained networks and

TABLE 3.1 Rubric for Rating Responses to Passage Retellings

Rating	Description
0	**Not rated.** Silent, irrelevant, or unintelligible.
1	**Minimal.**
2	**Limited.** Some concept sequences; missing major concepts and main narrative arc.
3	**Partial.** Several concept sequences and related parts of the main narrative.
4	**Adequate.** Enough major and minor concepts suggest main narrative logic.
5	**Good.** Major & minor concepts convey main narrative path and causal logic.
6	**Complete.** All major & many minor concepts support close narrative fidelity.

natural language processing. Specifically, Moby.Read uses Google's *word2vec* word vectors that were trained on about 100 billion words. The scoring model holds 300-dimensional vectors for 3 million words and phrases.

We collected human retelling judgments to validate the scoring model. Human judges rated retell responses on a 7-point scale shown in Table 3.1. Ratings from at least two different human raters were collected to each retell passage and each comprehension question. For comprehension questions, the human ratings provided models of answer quality spread over the scale range.

The retelling measured how semantically similar the words and sequences were between each passage and the student's retelling of that passage. These measures were combined over the three passages and then mapped onto a 0–4 scale as the retelling component of the comprehension score. For comprehension questions, similar techniques were used. The question component of the comprehension score was also mapped onto a 0–4 scale. Then, the mapped scores from retelling and from the comprehension questions were combined into an overall comprehension score.

Expression

Expression is the degree to which a student can clearly express the meaning and structures of the text through appropriate rhythm, phrasing, intonation, and emphasis. Expressive readings enhance the understanding and enjoyment of the text by a listener (Schwanenflugel & Benjamin, 2012). Scoring models for expression were based on human ratings. Raters were presented with rubrics and given training sets of responses so that they could practice rating actual student readings. Ratings were then compared with master ratings agreed upon by several assessment professionals with PhDs and at least 15 years of experience creating reading assessments. Discrepancies with the master ratings were reviewed and more training responses were presented to the rater until the rater was reliably assigning ratings consistent with the masters.

TABLE 3.2 Rubric for Expression	
Rating	**Description**
0	Insufficient sample for rating.
1	Word-by-word rendition with no reflection of word, phrase or sentence meaning.
2	Some local word grouping; little sentential phrasing.
3	Exhibits some text-inappropriate phrasing; sentence- and passage-level meaning is partially conveyed.
4	Prosody generally reflects meaning, but phrasing or intonation is sometimes inconsistent with the text.
5	Read for a listener; intonation, phrasing, and emphasis nicely express the meaning of the passage.

The data used to train the automatic scoring models had at least two human ratings of expression for each passage reading. Human raters used a 6-point scale to rate passage readings, as shown in Table 3.2.

Human judgments were then used to train a neural network with the goal of predicting how a human rater would rate the expression of the passage reading. The features used in the neural network were produced from the ASR system and included the pattern of phonetic segment durations and the log likelihoods of inter-word silence durations. The output values from the neural network were then mapped to a 5-level expression scale that ranges from 0 to 4.

VALIDATION

Machine scoring was validated in studies that compared the automated scores with human ratings. In addition, usability was measured through follow-up questions to evaluate ease of use and engagement of Moby.Read. We found that for several machine scores the correlation between the machine scores and the human ratings was better than our average human–human correlations. This is possible because the average human scores are generally much more reliable than individual human scores.

Study 1: Participants, Procedures, and Data Preparation

Participants

Participants in the study were 99 school-aged children from four different elementary schools: one public and one parochial school in New Jersey,

and two public elementary schools 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 second grade (29%), third grade (40%), and fourth grade (31%). With regard to ethnic background (using classifications set forth by the U.S. Census), 51 of the students were European American, 19 were African American, 4 were Asian American, and 25 were identified as Hispanic or Latino.

Procedure

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. In preparation for the experimental sessions, the facilitators set up two or three chairs about eight feet apart in a quiet area of the room or just outside the classroom. The Moby.Read assessment was delivered on an iPad Mini. Before the assessment, each student was fitted with a set of GearHead headphones with an inline microphone (the microphone was incorporated into the headset wire). 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.

For all graded items, responses were transcribed, readings were human-rated for expressiveness, and retells and comprehension questions were human-rated for comprehension. Following the administration, students were presented a brief usability survey.

Analysis

Five participants were screened out because their test responses were silent or completely unintelligible. This left 94 participants whose results are presented in these analyses.

Accurate Reading Rate

For accurate reading rate, two analyses were performed. The first was a comparison of accurate reading rate generated by the Moby.Read automated system versus scores generated from human transcriptions. For single passages, the correlation between these two scores was high, at $r = 0.96$.

Each recording of a passage reading was analyzed independently by two human raters. The human raters had PhDs and at least 10 years of

experience in reading assessment. Each rater listened to recorded student reading, marked errors, and measured the length of time during which the student read. Raters computed the WCPM for the passage by dividing the number of words read correctly over the duration of time of the reading (in seconds) times 60 (to place the units in minutes). Inter-rater reliability in this task was 0.99. Averages of the human-computed WCPM values were computed and the median score for each 3-passage session was derived for each participant. These median human scores were then correlated with the median machine-generated WCPM score. The resulting correlation coefficient was 0.99, confirming that the median scores produced by the Moby.Read app are comparable to median value of scores produced by human raters. Figure 3.7 presents a scatterplot of machine versus human scores at the session level. This comparison of median WCPM between the two methods suggests that scoring based on automatic speech processing and alignment with text has a high degree of accuracy.

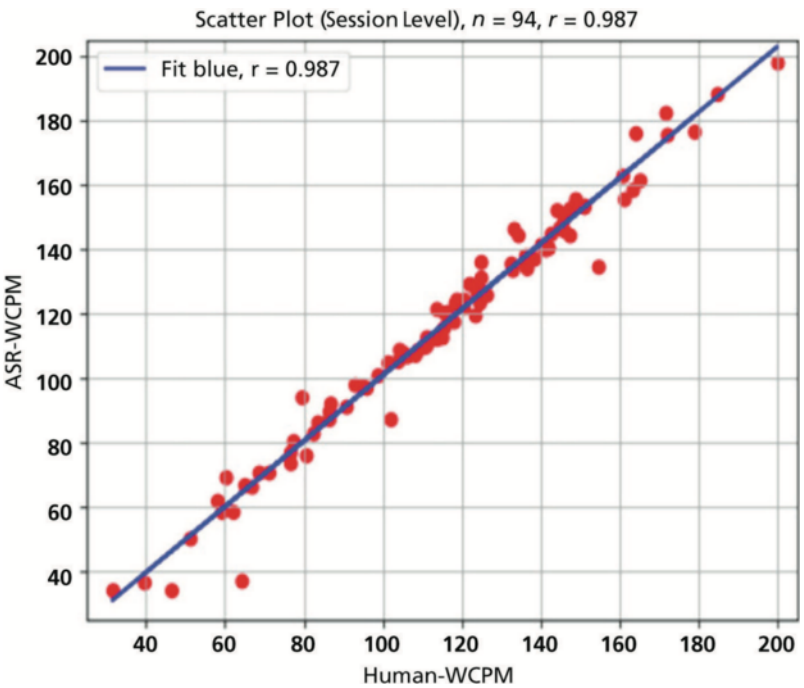

Figure 3.7 Accurate Reading Rate (WCPM) scores from human judges vs. machine (ASR).

Comprehension

We validated the comprehension scores from passage retelling. Machine-generated comprehension scores on retellings were compared with human ratings of retellings. Average machine scores were generated for each participant and were correlated with average human ratings. The human–machine correlation coefficient was 0.92. This correlation was better than that of the average human–human inter-rater correlation of 0.88. A scatterplot of the scores is presented in Figure 3.8.

Expression

To verify expressiveness scores, Moby.Read scores were compared to an average of three human ratings of expressiveness. For the three pairs of human raters, the average inter-rater correlation at the response level was 0.74. The correlation coefficient of machine-generated expressiveness scores and average human ratings of expressiveness was 0.88 (0.94 at the

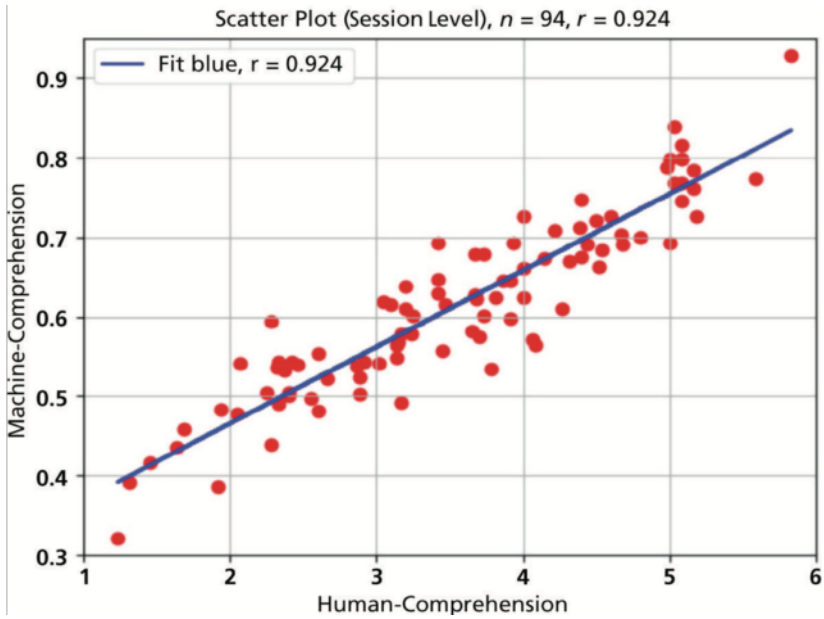

Figure 3.8 Plot of Comprehension scores from human raters and from machine models.

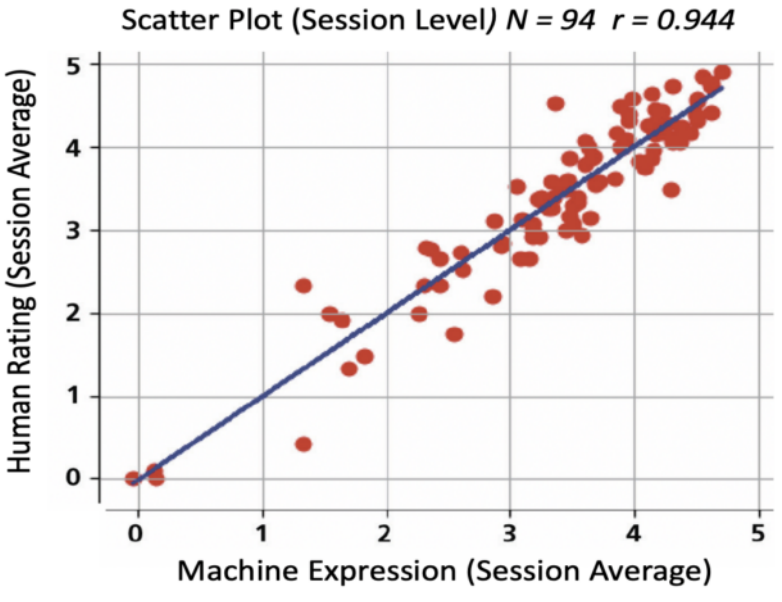

Figure 3.9 Plot of Expression scores from human judges vs. machines.

session level; see Figure 3.9), a statistically significant improvement. These correlations indicate that the machine scores were more reliable in producing consistent expressiveness scores than human raters.

Usability

Usability was evaluated on the iPad device at the end of the Moby.Read assessment. Students were presented the image shown in Figure 3.10 and

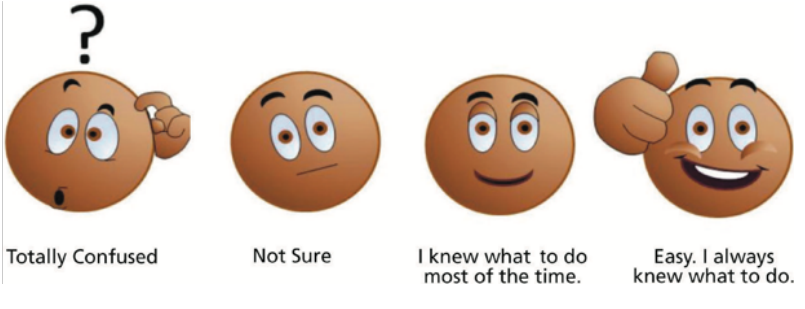

Figure 3.10 Four point rubric displayed on screen to students.

asked to tap the screen image that best represented their experience with the app.

Two students failed to respond to any items. Among the remaining 97 who responded, 48 selected *Easy. I always knew what to do*; 42 selected *I knew what to do most of the time*; and 7 selected *Not sure*. No students selected *totally confused*. These results show that 97 of 99 (or 98%) students knew what to do most or all of the time, suggesting self-administration is viable for the majority of students.

Study 2: Concurrent Validity

A second study compared the scoring derived from Moby.Read with that of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) NEXT assessment (Dynamic Measurement Group, 2018), a widely used human-administered and scored ORF assessment.

Participants

Twenty students from an elementary school in California participated. Nine were female and 11 were male. Seven were in second grade, six were in third grade, and seven were in fourth grade. Students were given a $15 gift card to a local toy store as remuneration for their participation.

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 a 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 portion, each child was fitted with a GearHead microphone headset and the Moby.Read test was self-administered and automatically scored by the Moby.Read system. Students were administered a form appropriate for their grade. Moby.Read was delivered on an iPad Mini.

For the DIBELS assessment, students were administered the fall benchmark test form, which consisted of three grade-level passages of about 250 words in length. The administrator followed the administration and scoring procedures described in the test's official documentation (Good et al., 2011). Students were given a passage and asked to read it out loud using the instruction prompts from the DIBELS assessment manual. The administrator started a timer when the student started reading the first word of the passage. As the student read the passage, the administrator marked reading errors on a scoring sheet. After one minute, the timer beeped and the student's place in the passage was marked on the scoring sheet (none

of the students finished the passage). Then the student was asked to tell the administrator about the story. Responses were timed for 1 minute and marked on a comprehension scoring sheet which tracked the number of words spoken in the student's response.

At the conclusion of the session, the administrator pointed out that the student had done both a test on the iPad ("Moby.Read") and on paper ("teacher administered") and asked which the students preferred, and why. Consistent with standard practice, after the session, the administrator used the scoring sheets to calculate errors and WCPM.

Concurrent Results

Moby.Read accurate reading rate scores from 20 participants were compared to scores from the ORF task of DIBELS NEXT. The correlation between the two scores was 0.88. Published studies investigating DIBELS report a test-retest reliability of 0.82 and an inter-rater reliability of 0.85 (Goffreda & DiPerna, 2010). The reliability of an instrument limits the strength of the correlation between that instrument and others measuring the same construct. So, the correlation with Moby.Read is at the ceiling of what would be expected given the reliability of DIBELS.

When asked which experience they preferred, 18 out of 20 students (90%) said they preferred "Moby.Read," and 2 students said they preferred "both"; no students preferred the teacher administration. Useful qualitative information was provided by the students when asked why they preferred the Moby.Read administration. The feedback can be clustered as shown in Table 3.3.

TABLE 3.3 Examples of Student Responses When Asked Why They Preferred the Automated Assessment to the Human Administered Test

Theme	Student Feedback
Technology	• "I like the screens" • "There's this Siri thing"
Questions	• "It asks questions, so you're reading for purpose" • "You get to answer questions"
Re-read option	• "I can read the stories again" • "You could read it again"
Administration	• "The iPad tells you what to do" • "It tells you about the story before you read it"
Privacy	• "It's more private" (i.e., no teachers are watching and judging)
User interface	• "You have more time so you can finish the story" • "More pictures and stuff"

These qualitative data provide support for an assertion that this tablet-enabled assessment that allows students to self-administer ORF can produce an engaging experience for students.

DISCUSSION OF MOBY.READ

Several forms of evidence support the validity and utility of the Moby.Read instrument. As a voice-interactive digital app, students found the assessment engaging and easy to use. The Moby.Read test structure and instruction format have been developed iteratively through user studies and feedback. The pilot studies demonstrate that the Moby.Read test experience was sufficiently self-explanatory for 98% of students to successfully self-administer the tests without teacher intervention. Further, 90% of students indicated they enjoyed using the Moby.Read application more than traditional ORF as administered by a teacher.

The Moby.Read app also includes ready tools for teachers to monitor student progress over time and to share recordings of a student's performance with parents or with a reading specialist. Two of the teacher's displays are shown in Figure 3.11.

Figure 3.11 Two displays for teachers. Left supports playing responses; right tracks progress.

In summary, a Moby.Read assessment measures the three components of ORF—accuracy, accurate rate, and expression—while providing a concurrent measure of reading comprehension that helps locate a student's reading level. Student data indicates that these scores are comparable to scores produced by human raters. In some cases, such as for comprehension and expression, the scores produced by Moby.Read are closer to average human ratings than scores from individual human raters are to each other. Empirical data also provide evidence of construct validity with a high correlation between Moby.Read scores and fluency scores from other standardized reading assessments. Finally, the content in the Moby.Read passages samples a broad range of texts appropriate for students in Grades 1 through 5. In sum, evidence supports the use of Moby.Read scores as a reliable measure of ORF and indicator of reader level.

CONCLUSION

For assessing spoken language proficiency and oral reading, advances in ML have brought computers the ability to present spoken items and score spoken responses in ways that match or exceed the accuracy of single human scorers. As the technology advances and more individuals acquire skill in ML, the costs of creating such systems is diminishing. Recognizing the trend, school districts and state departments of education are increasingly evaluating automated scoring as an option (e.g., Arizona Department of Education, 2013; Texas Education Agency, 2018). Evanini, Hauck, and Hakuta (2017) provide a practical list of considerations for schools or organizations intending to use AI-based approaches in assessing language.

Fundamentally, the studies reported here demonstrate that thoughtful application of ML to assessment enables improved score reliability (compared to human judgments). AI scoring can also promote the use of preferred assessment methods (e.g., comprehension measured from passage retelling) that previously have been too time-consuming, expensive, or unreliable for general use. Newer assessments that use AI scoring can provide rich and timely information to teachers, reading specialists, and administrators, in ways that relieve their assessment burden while improving accuracy and reliability.

REFERENCES

Arizona Department of Education. (2018). Assessment: Azella. Retrieved from https://www.azed.gov/assessment/azella/

Bernstein, J. (2012). Computer scoring of spoken responses. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics.* https://doi.org/10.1002/9781405198431.wbeal1044

Bernstein, J., & Franco, H. (1996). Speech recognition by computer. In N. Lass (Ed.), *Principles of experimental phonetics* (pp. 408–434). St. Louis, MO: Mosby.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355–377.

Cheng, J., Zhao D'Antilio, Y., Chen, X., & Bernstein, J. (2014, June). Automatic spoken assessment of young English language learners. Proceedings Ninth Workshop on Innovative Use of NLP for Building Educational Applications, Baltimore, MD.

Common Core. (2011). *Learning less: Public school teachers describe a narrowing curriculum.* Retrieved from https://www.americansforthearts.org/sites/default/files/cc-learning-less-mar12.pdf

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.* Retrieved from http://www.corestandards.org/assets/E0813_Appendix_A_New_Research_on_Text_Complexity.pdf

Daane, M., Campbell, J., Grigg, W., Goodman, M., & Oranje, A. (2005). *Fourth-grade students reading aloud: NAEP 2002 special Ssudy of rral reading* (NCES 2006-469). U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. Washington, DC: Government Printing Office.

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

Dynamic Measurement Group. (2018). *DIBELS.* Retrieved from https://dibels.org/dibels.html

Evanini, K., Hauck, M. C., & Hakuta, K. (2017). Approaches to automated scoring of speaking for K–12 English language proficiency assessments. *ETS Research Report Series, 2017,* 1–11. https://doi.org/10.1002/ets2.12147

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.

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.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.) New York, NY: Springer.

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

Houghton Mifflin Harcourt. (2018). *EasyCBM.* Retrieved from https://www.easyCBM.com

Morrow, L. M. (1985). Retelling stories: A strategy for improving young children's comprehension, concept of story structure, and oral language complexity. *The Elementary School Journal, 85*(5), 647–661.

National Reading Panel (U.S.), & National Institute of Child Health and Human Development (U.S.). (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.* Washington, DC: National Institute of Child Health and Human Development, National Institutes of Health.

Pearson. (2018a). *aimswebPlus.* Retrieved from https://www.aimswebplus.com/

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Vesely, K. (2011). The KALDI speech recognition toolkit. In *Proceedings of IEEE 2011 workshop on automatic speech recognition and understanding.* Big Island, HI: IEEE Signal Processing Society.

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.

Schwanenflugel, P. J., Hamilton, A. M., Kuhn, M. R., Wisenbaker, J., & Stahl, S. A. (2004). Becoming a fluent reader: Reading skill and prosodic features in the oral reading of young readers. *Journal of Educational Psychology, 96,* 119–129.

Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. C. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21,* 459–479.

Texas Education Agency. (2020). *TTELPAS resources.* Retrieved from https://tea.texas.gov/student-testing-and-accountability/testing/texas-english-language-proficiency-assessment-system-4

Van Moere, A., & Downey, R. (2016). Technology and artificial intelligence in language assessment. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 341–358). Boston, MA: Walter de Gruyter.

Wilson, R. M., Gambrell, L. B., & Pfeiffer, W. R. (1985). The effects of retelling upon reading comprehension and recall of text information. *The Journal of Educational Research, 78*(4), 216–220.

Young, S. (1996). Large vocabulary continuous speech recognition. *IEEE Signal Processing Magazine, 13*(5), 45–57.

Zhang, X., Trmal, J., Povey, D., & Khudanpur, S. (2014). Improving deep neural network acoustic models using generalized maxout networks. Retrieved from https://www.danielpovey.com/files/2014_icassp_dnn.pdf