Validation of Automated Scoring of Oral Reading

Article *in* Educational and Psychological Measurement · June 2012

DOI: 10.1177/0013164411412590

6 authors, including:

Jared Bernstein
Stanford University
88 PUBLICATIONS **1,071** CITATIONS
SEE PROFILE

Alistair Van Moere
13 PUBLICATIONS **258** CITATIONS
SEE PROFILE

Masanori Suzuki
Analytic Measures Inc
12 PUBLICATIONS **82** CITATIONS
SEE PROFILE

Educational and Psychological
Measurement
72(3) 435–452
© The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164411412590
http://epm.sagepub.com

Validation of Automated Scoring of Oral Reading

Jennifer Balogh[1], Jared Bernstein[1],
Jian Cheng[1], Alistair Van Moere[1],
Brent Townshend[1], and
Masanori Suzuki[1]

Abstract

A two-part experiment is presented that validates a new measurement tool for scoring oral reading ability. Data collected by the U.S. government in a large-scale literacy assessment of adults were analyzed by a system called VersaReader that uses automatic speech recognition and speech processing technologies to score oral reading fluency. In the first part of the experiment, human raters rated oral reading performances to establish a criterion measure for comparisons with the machine scores. The goal was to measure the reliability of ratings from human raters and to determine whether or not the human raters biased their ratings in favor of or against three groups of readers: Spanish speakers, African Americans, and all other native English speakers. The result of the experiment showed that ratings from skilled human raters were extremely reliable. In addition, there was no observed scoring bias for human raters. The second part of the experiment was designed to compare the criterion human ratings with scores generated by VersaReader. Correlations between VersaReader scores and human ratings approached unity. Using G-Theory, the results showed that machine scores were almost identical to scores from human raters. Finally, the results revealed no bias in the machine scores. Implications for large-scale assessments are discussed.

Keywords

automated scoring, oral reading fluency, speech recognition

[1] Pearson, Knowledge Technologies, Group, Palo Alto, CA, USA

Corresponding Author:
Alistair Van Moere, Pearson, Knowledge Technologies Group, 299 S. California Avenue, Suite 300,
Palo Alto, CA 94306, USA
Email: Alistair.VanMoere@pearsonkt.com

Background

Oral reading fluency has been identified by the National Reading Panel as one of five key areas of focus in reading instruction (Armbruster, Lehr, & Osborn, 2001). Even so, oral reading fluency is absent from most large-scale reading assessments. For example, the National Assessment of Educational Progress (NAEP), a large-scale assessment administered to more than 190,000 fourth-graders and more than 160,000 eighth-graders in the United States in 2007, provides important information about national trends over time but does not include an assessment of oral reading fluency. In a separate study conducted by NAEP in 2002, oral reading fluency was analyzed for 1,779 fourth graders (Daane, Campbell, Grigg, Goodman, & Oranje, 2005). The much smaller sample size for this study compared with the full-scale NAEP assessment (less than 1%) reflects the logistical challenges inherent in collecting the data for an oral reading fluency study and having trained expert raters listen to each response and score it. Because the scoring task requires hands-on expertise, it can become time consuming and expensive. These problems may be compounded with the issues of reliability. Procedures must be adopted to ensure that raters are consistent within their own ratings and that ratings from one rater do not differ significantly from ratings of other raters.

A possible solution to the problem of inefficient and costly oral reading fluency assessments is the development of a computerized measurement tool. One approach is to apply speech processing technologies, including automatic speech recognition, to create an automated assessment. Such a tool, called VersaReader (www.VersaReader.com), was developed by Ordinate Corporation, which is now a part of Pearson's Knowledge Technologies group.

A basic question when considering automated measurement of oral reading is whether or not the automated technologies can score reading responses accurately. The goal of automated scoring is to produce an end product, such as the number of words read correctly, that is similar to what expert human raters would produce. Ideally, the machine scores match human ratings within an acceptable range and are as reliable if not more reliable than the human ratings. Previous research results with other speech technologies that assess oral reading performance have approached these targets, but still automated scoring has never quite reached a level as accurate as the average human rater (Adams, 2006; Black, Tepperman, Lee, & Narayanan, 2008; Mostow, Roth, Hauptmann, & Kane, 1994; Mostow, Aist, Huang, Junker, Kennedy, Lan, et al., 2008). For large-scale assessments that potentially influence policy decisions, it is critical that the machine scores and human ratings are comparable. An opportunity to examine the question of machine accuracy was presented when the VersaReader technology was used to digitize and score oral reading performances in the National Assessment of Adult Literacy, a large-scale literacy assessment.

Fluency Addition to the National Assessment of Adult Literacy

In 2003, the National Center for Education Statistics (NCES) of the U.S. Department of Education administered its continuing assessment of English language literacy skills of American adults called the National Assessment of Adult Literacy (NAAL). The purpose of the large-scale assessment was to characterize the status of adult literacy in the United States, report on national trends, and identify relationships between literacy and selected characteristics of adults. Based on a recommendation from a panel of experts, the 2003 administration of the assessment included oral reading fluency. The goal of the Fluency Addition to NAAL (FAN) was to complement the functional literacy focus of the main NAAL study and measure a reader's basic reading skills through analysis of oral reading performance. One of the perceived benefits of the FAN was to better understand the basic reading skills of adults reading at the lowest levels. Given that more than 100,000 oral reading responses had to be scored, VersaReader was used to automate and augment the analysis of oral readings.

This article describes a two-part experiment conducted to validate scores generated by VersaReader. The VersaReader system uses speech processing technologies to produce scores of oral reading fluency automatically. Because VersaReader can score a 1-minute oral reading performance in a few seconds and does not require trained expert raters, it shows promise as a solution to certain logistical problems in large-scale reading assessments. In the first part of the experiment, human raters rated a subset of oral reading performances collected from the FAN. The goal was to produce reliable and unbiased ratings that would serve as a criterion for machine-generated scores. As such, the experiment was designed to measure the reliability of human ratings and to determine whether or not the human raters biased their ratings in favor of or against three groups of readers: Spanish speakers, African Americans, and all other English speakers. In the second part of the experiment, the human ratings were compared with VersaReader scores. The validation experiment provides evidence that a machine can score oral reading as well as expert human raters do. In addition, it presents data for oral reading assessment of adults as opposed to children.

Experiment 1A

The goal of the first part of the experiment was to establish criterion scores from expert human raters that could be used for comparisons with machine-generated scores. The experiment was designed to answer two questions: (a) How reliable are ratings from human raters? and (b) Do these human ratings exhibit any bias? Specifically, do human raters associated with a linguistic/ethnic group give differentially higher or lower ratings to readers in the same linguistic/ethnic group and/or different linguistic/ethnic groups? For this experiment, three linguistic/ethnic groups were identified:

- Native Spanish speakers (SP)
- Native English-speaking African Americans (AA)
- Other native English speakers (OE)

These three groups were selected for the validation experiment for several reasons. First, the members of the SP group and the AA group were thought by NCES to be most often subjected to negative bias. Furthermore, with regard to speaking patterns, the OE group is often associated with "standard" linguistic forms, whereas members of the other two groups are not. Finally, the three groups were relatively well sampled in NAAL, providing a sufficient amount of data to perform the analyses.

Method

Fluency Addition to NAAL.

FAN participants. The sample of participants who provided oral readings for the FAN were from the 2003 NAAL data collection and represented the demographic profile of the United States. Three percent of the FAN participants in the data collection could not complete the easiest items of the main NAAL assessment and were administered an Adult Literacy Supplemental Assessment (ALSA) instead. The ALSA was more appropriate for participants at a very low reading level. In the ALSA, participants were asked to read things such as the text on a grocery advertisement or television program schedule. Both the participants who completed the main NAAL and those that took the ALSA were given the FAN. The NAAL study also included a sample of participants who were currently in prison.

The number of participants with completed interviews in the NAAL data collection including the ALSA and prison sample totaled 19,258. From this sample, responses from 480 participants who took either the main NAAL or the ALSA were sampled for the validation experiment. The prison data collection was still ongoing at the time of the experiment; therefore, the prison sample data set was not included in the analysis.

Of the 480 participants, 160 were from each of the three linguistic/ethnic groups: SP, AA, and OE. The 480 participants were selected at random from their respective demographic groups within deciles of the whole NAAL population sample. First, the entire NAAL/ALSA sample was stratified into 10 bins based on the rank score on the main NAAL assessment, which presented a series of literacy tasks designed to measure prose literacy, document literacy, and quantitative literacy. Sixteen participants from each linguistic/ethnic group (SP, AA, and OE) were sampled from each of the 10 bins for the validation experiment. For example, the first bin consisted of all the participants who scored in the 0 to 10th percentile. From this bin, 16 SP participants, 16 AA participants, and 16 OE participants were sampled. The second bin contained participants who scored in the 11th to 20th percentile. Sixteen participants from each linguistic/ethnic group were selected from this bin, and so on. This approach to stratification was used so that readers with similar ranges of abilities from each of the three linguistic/ethnic groups were sampled for the experiment.

FAN procedure. In the FAN administration, each participant was presented with 10 texts to read: a digit list, a letter list, three word lists, three pseudoword lists, and two passages of about 150 to 200 words in length. One passage was at an elementary school level of difficulty, whereas the other was at a middle school level of difficulty. After reading each passage, the participant was asked a brief comprehension question, for example, “What is the age-old remedy for a cold?” The comprehension question was not scored.

During a session, a participant was fitted with a set of headphones and a microphone that connected to a laptop computer. The administrator viewed the laptop screen during the session, while the participant read from printed materials. For each task, the participant was given a fixed time window in which to read the material. For lists (digits, letters, words, and pseudowords), the fixed time window was 20 seconds, and for passages, the time window was 60 seconds per passage. Each list and passage were digitized and saved to a separate audio file referred to as a response recording.

Details of the participant sample, reading materials and procedure for the FAN are available in a public report by the National Center for Education Statistics (Baldi, 2009).

Human Rater Participants. Ten human raters active in reading education participated in the experiment. Seven had advanced degrees in education or language-related fields including Curriculum and Instruction, Education, Educational Leadership, Foreign Language Education, Linguistics, Reading, and Teaching English to Speakers of Other Languages. Of the three without graduate degrees, two held teaching certificates and one was enrolled in a teaching certification program for reading. All had extensive experience teaching reading.

Raters were assigned to one of the three linguistic/ethnic groups (SP, AA, or OE). Classification was based on the rater’s own linguistic/ethnic group and/or the rater’s extensive experience teaching students of one of these groups. Of the 10 raters, three were classified as SP, four as AA, and three as OE.

The three Spanish raters took Knowledge Technologies’ Versant™ tests for both Spanish and English (www.VersantTest.com) to verify that they spoke fluent Spanish. Both tests generate an overall score ranging from 20 to 80. Native speakers score at the very high end of the scale (e.g., 98% of native Spanish speakers score more than 70 on the Versant Spanish test). The SP raters scored 74, 75, and 80 on the Versant Spanish test. All three SP raters received a score of 80 on the Versant English test.

Experimental Design. Two reading responses from each of the 480 participants were analyzed. For 50% of the participants (240), two different passages were analyzed in the study, for 25% (120 participants) two word lists were analyzed, and for the remaining 25% (120 participants) one passage and one word list were analyzed. A greater number of passage pairs were included since there were eight different passages and only three different word lists presented in the NAAL data collection.

The design of the validation experiment was factorial with respect to rater and participant linguistic/ethnic groups. Each of the two response recordings from each participant was judged by at least one rater from each linguistic/ethnic group. This produced 2,880 scores (2 response recordings × 480 participants × 3 rater groups).

In addition, at least 20% of the responses were rated by more than one rater in the same linguistic/ethnic group. The two ratings from within the same rater group were used in intragroup reliability analyses. Moreover, 5% of the material was randomly presented to the same rater twice to measure within-rater variability.

Method and Apparatus for Human Rating.

The framework for the human rating of oral reading was based on the Qualitative Reading Inventory-3 (Leslie & Caldwell, 2001), a traditional oral reading assessment method. Since the goal of the validation study was to compare VersaReader with an independent, well-understood process of human rating, it was vital to keep to known practices. The goal for NAAL was not to develop an improved method of human rating of reading but rather to use an example of current good practice to evaluate a new machine-scoring approach. The Qualitative Reading Inventory-3 served this purpose well because it is a widely used inventory representative of common practices in human-judged reading assessment with well-described procedures. In the Qualitative Reading Inventory-3, human raters classify errors according to whether they are substitutions, insertions, omissions, reversals, or self-corrections.

The human raters listened to and entered ratings for the response recordings using a web-based rating system. The rating system semirandomly selected response recordings for each rater to score. The selection algorithm was designed to assign each response recording to a member of each rater group at least once. In addition, 20% of the time, the system presented raters with response recordings that had already been rated by a member of the same group, and 5% of the time, it selected a response recording that had already been rated by the same rater. Other than these constraints, the response recordings were presented to raters randomly, with a different order for each rater. Each rater worked independently from the other raters.

The rating system displayed the source text for the passage or word list without punctuation. Response recordings were played from a cached audio file so that there would be no artificial breaks in the playback because of Internet connectivity problems. The rater could repeat any section of the response by moving an audio scroll bar.

Raters were also provided with paper scoring sheets that presented the original source text including slashes for line breaks and an index number under each word.

Procedure.

After raters were selected to participate in the experiment, they received a *Rater Instruction Manual* that described how to access and listen to response recordings, evaluate reading errors, and enter ratings into the web-based rating system. The following criteria were used by human raters to determine if a word was read accurately:

- The word was produced as commonly pronounced in any functional English-speaking pattern, and speaking pattern was generally evident in the rest of the reader’s response.
- The word as pronounced was not closer to the pronunciation of another known word in that speaking pattern.
- There was not an intrusive substitution or deletion of segments that indicated that the word was not understood or recognized by the reader.

The term *speaking pattern* included patterns associated with regional dialects; social dialects; commonly encountered dysarthrias, such as lisping; and nonnative varieties of English learned as a second language. The key qualifier was that the speaking pattern be in functional use among some identifiable group of people.

Raters were given 12 practice response recordings to analyze. The same training materials were analyzed by all raters. (The response recordings used for training purposes were not evaluated in the experiment.) The response recordings were from participants with various levels of reading ability and from various linguistic/ethnic backgrounds.

Raters were given a paper scoring sheet and were asked to box the index number of the first word attempted and the index number of the last word attempted in the passage. They were also instructed to mark all reading errors.

After raters had finished the first set of practice responses, their ratings were discussed with the experimenter. Once the experimenter was confident that each rater could perform the task, the raters were sent a packet of paper scoring sheets for eight passages and three word lists and a PIN granting the rater access to the rating system.

During the rating task, raters first marked errors for a given response on a paper scoring sheet and then entered scores for that response into the rating system.

The following equations were used to calculate the scores for passages:

$$\begin{aligned}\text{Total Words Attempted} &= \text{Index number of the last word attempted} \\ &\quad - \text{Index number of the first word attempted} + 1 \qquad (1)\end{aligned}$$

$$\begin{aligned}\text{Number of Words Read Correctly} &= \text{Number of Words Attempted} \\ &\quad - \text{Substitutions} - \text{Omissions} - \text{Insertions} - \text{Reversals} \qquad (2)\end{aligned}$$

Knowledge Technologies reasoned that although *self-corrections* are reading errors, they should not be deducted from the *number of words read correctly* since participants did eventually read these words and phrases accurately.

All scoring sheets were returned to Knowledge Technologies and verified. Faulty database entries due to arithmetic errors, counting errors and premature data submissions were corrected.

Results

Because the rating system served responses to raters in random orders, the number of items scored by each rater group was slightly less than that described in the design. Specifically, there were 12 fewer passages scored (five fewer from the SP raters, one from the AA raters, and six from the OE raters) and four fewer word lists scored (two fewer from the SP raters and two from the OE raters). Also, one response was excluded from the analysis because it was used as a training item.

Table 1. Intrarater Reliability Coefficients

Rater	Count	Reliability coefficient
SP-1	25	1.00
SP-2	25	1.00
SP-3	24	1.00
AA-1	26	1.00
AA-2	56	1.00
AA-3	23	1.00
AA-4	24	1.00
OE-1	23	1.00
OE-2	25	0.99
OE-3	28	1.00
Mean	28	1.00

Table 2. Intragroup Reliability Coefficients for the Three Rater Groups

	Intragroup reliability coefficients			
Rater group	Count	Passages	Count	Word lists
SP	153	1.00	91	0.98
AA	514	0.99	331	0.99
OE	156	1.00	91	1.00
Mean		1.00		0.99

Note: SP = Native Spanish speakers; AA = Native English-speaking African Americans; OE = Other native English speakers.

Reliability

Intrarater reliability of ratings. Within-rater reliability of ratings was estimated by calculating the correlation coefficient of the items scored by the same rater twice. Table 1 shows the coefficient for each rater's ratings. Because of the small number of data points, data for passages and word lists were combined. The intrarater reliability coefficients were near perfect.

Intragroup reliability. The intragroup reliability coefficients were calculated by correlating ratings of the same item from more than one member of the same rater group. The intragroup reliability coefficients for the three groups are listed in Table 2.

Note that the number of responses rated by two different members of the AA group is quite large. The reason for this is that there were four raters in this group instead of three; therefore the frequency of rating a response that had already been rated by another member of the AA group was high.

The average intragroup reliability coefficient for the number of words read correctly is also near perfect for both passages and word lists.

Human Rater Bias. The second set of analyses addressed whether or not the human raters introduced a bias with regard to the participants' linguistic/ethnic classification.

Table 3. Performance Scores for Passage Readings for Each Linguistic/Ethnic Group

		Number of words read correctly		
Participant group	Count	Mean	Standard deviation	Standard error
SP				
SP raters	187	115	42	3.1
AA raters	190	117	42	3.0
OE raters	188	115	42	3.1
AA				
SP raters	189	121	44	3.2
AA raters	190	122	45	3.3
OE raters	187	122	44	3.2
OE				
SP raters	193	132	41	3.0
AA raters	193	132	41	3.0
OE raters	193	132	41	3.0

Note: SP = Native Spanish speakers; AA = Native English-speaking African Americans; OE = Other English speakers.

First, descriptive statistics for each rater and participant group were generated. The means and standard deviations of the human ratings for each of the rater and the participant groups for passage readings are presented in Table 3. Scores of 0 because of wrong passage readings (7 responses) and no audio (19 responses) were excluded from the analysis. Also one response was excluded from the analysis because it was used as a training item.

To determine whether or not there were statistically significant main effects of participant group, rater group, and/or an interaction between the two, a two-way analysis of variance (ANOVA) was used to analyze the data. The two factors were Participant Group (SP, AA, and OE) × Rater Group (SP, AA, and OE). Alpha for all statistical tests was .05.

The ANOVAs revealed a statistically significant main effect of participant group, $F(2, 1701) = 20.9$, $MS_{\text{Participant}} = 37{,}770$, $p < .01$. This main effect merely demonstrates that the overall reading performance scores of the three participant groups were different, despite the stratified sampling.

The main effect of rater group was not statistically significant, $F(2, 1701) = 0.07$, $MS_{\text{Rater}} = 120$, $p = .94$. This finding suggests that there were no statistically significant differences between the ratings from the three rater groups.

There was also no statistically significant interaction, $F(4, 1701) = 0.02$, $MS_{\text{Participant Rater}} = 41$, $p = .99$, which indicates that no scoring bias was detected in the human ratings.

The patterns for word lists were similar. ANOVAs for word lists revealed a statistically significant main effect of participant group, $F(2, 1049) = 26.7$, $MS_{\text{Participant}} = 1{,}801$, $p < .01$, no main effect of rater group, $F(2, 1049) = 0.09$, $MS_{\text{Rater}} = 6.2$, $p = .91$, and no interaction, $F(4, 1049) = 0.07$, $MS_{\text{Participant Rater}} = 4.6$, $p = .99$.

Discussion

The results show that the average intrarater and intragroup reliability coefficients for ratings were near perfect. The intrarater reliability coefficients indicate that the human raters were consistent with their own scores. The intragroup coefficients show that the ratings from raters within each linguistic/ethnic group were consistent with other ratings within the same group.

The ANOVAs revealed a main effect for participant group for both passages and word lists. This main effect merely demonstrates that the overall reading performance scores of the three participant groups were different. That is, the African American group scored higher than the group of native Spanish speakers, and the Other English speaking group scored higher than the African American group. The result is somewhat unexpected given the parallel decile sampling. Even though each group's average scores (after screening) were about the same on the main NAAL assessment (SP = 49, AA = 47, and OE = 48), the participant groups were statistically different with regard to oral reading ability.

The results of the analyses showed no statistically significant main effect of rater group. This finding suggests that there were no differences between the ratings from the three rater groups. The SP, AA, and OE raters all rated the oral reading responses in the same way, producing almost identical scores. There was also no statistically significant interaction, which indicates that no scoring bias was detected in the human ratings.

Reliable and unbiased ratings generated by human experts are a prerequisite for the ratings to act as a criterion measure to evaluate machine scores. Taken together, the results suggest that the human ratings used as reference criteria for machine score comparisons were reliable and unbiased.

Experiment 1B

The purpose of the second part of the validation experiment was to provide evidence that machine-generated scores were comparable to the human ratings and also to show that the machine-generated measures did not introduce bias with respect to participants' linguistic/ethnic classification.

Method

Materials

The same response recordings collected from the FAN and sampled for Experiment 1A were analyzed in the second part of the experiment.

Procedure for Developing VersaReader Automated Scoring. Although speech recognition systems are similar in many ways, the technology in VersaReader was customized for scoring the FAN responses. To accommodate foreign accents, acoustic models (representing the sounds or phones produced when speaking English) were trained from speech of nonnative speakers. The VersaReader dictionary lists the most common pronunciations for each word that the system should recognize. Every word that appeared in the FAN materials was entered into the system's dictionary. In addition, entries were created for common substitutions of words in the source text, based on FAN data.

The language model is a representation of the sequence of words the speaker is expected to say. The language models not only contain the most likely strings of words that a reader is expected to say but also the types of mistakes and disfluencies that readers are most likely to make. The reading errors can be represented as a list of "rules" (X → Y) with a probability associated with each one. For example, if the printed word is *a*, and readers commonly say the word *the*, the rule for this reading error would be "*a* → *the*" (Cheng & Townshend, 2009).

Two sets of data were created: a training set of 4,681 responses and a test set of 2,170 responses. The training set was used to build a language model rule set for each passage and word list and the test set was used to test the models. The two sets did not intersect. Because a dense sample of noncanonical readers was required to build the FAN language model rule sets and because most responses were without error, Knowledge Technologies oversampled respondents who scored at the lower end of the performance scale on the general NAAL assessment.

The VersaReader speech recognition system formulated a hypothesis of what the speaker said and identified the string of words that best matched the participant's speech. This hypothesis was compared with the source text. Using a standard string alignment algorithm that minimizes the sum of word deletions, substitutions, and insertions, the participant's response was aligned with the source text. The reading errors were then tallied and weighted, and a final value of the number of words read correctly was generated. For participants with very little reading ability who took the ALSA instead of the main NAAL, all the response recordings were hand transcribed and then automatically aligned with the human transcriptions. Thus, instead of speech recognition, automatic alignment of hand transcriptions was done to improve accuracy of the ALSA scores for the entire FAN sample delivered to NCES. Although this step removes some automation from these scores, the automated alignment algorithm was still applied. Other information was also extracted from the participant's utterance such as the duration of speech, the rate of speech, and pause duration.

Results

Comparison of Machine Scores and Human Ratings. The first set of analyses directly correlated human ratings with VersaReader scores. The human–machine correlations were evaluated against human–human correlations in which a single rater's score was compared with the average of the scores from all the other raters who rated the response recording. Single-rater scores were correlated with average scores so that the human–human comparison would be similar to human–machine correlations in which a single machine score is compared with the average of the human ratings.

Table 4. Correlation Coefficients of the Average Human Rating and One Human Rating With the Average Human Rating and the Machine

	Response count[a]	Average human–one human	Average human–machine
Same items			
Passage–passage	479	1.00	0.99
Word list–word list	240	1.00	0.99
Passage–word list	240	1.00	0.99

[a] One response was excluded from the passage–passage analysis because it was used as a training item.

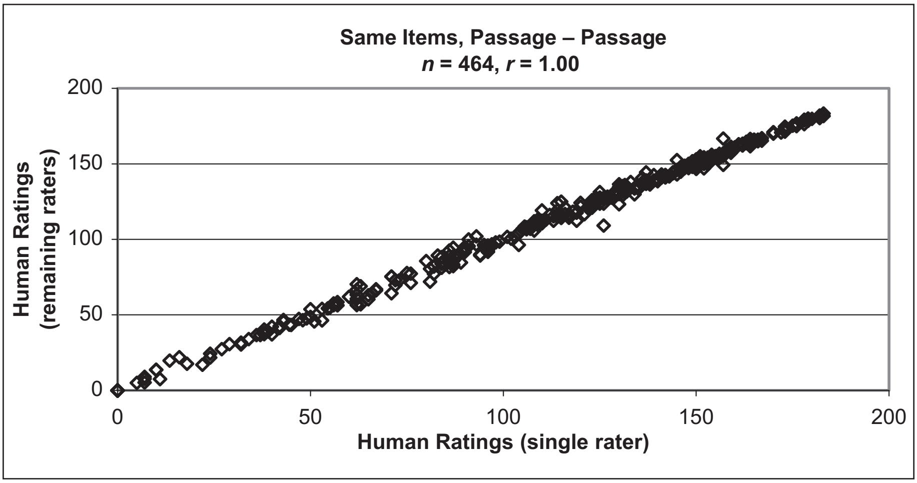

Figure 1. Human versus human scatter plot for number of words read correctly
Note. There were eight respondents in the passage–passage group for whom there were not enough ratings across the different raters to do the human–human analysis, so their 16 responses were excluded. The response that was used as a training item was one of the excluded responses.

To estimate the correlation of a single rater versus all the other raters, random samples were drawn such that (a) scores from different individual raters were used for different items and (b) each rater provided scores for about an equal number of items. The random sampling was done five times, and the correlations were averaged together to produce average correlation values of one rater with the average of the scores from the remaining rates who scored that item. For the human–machine correlations, an average of all available human rater scores was computed (from three to seven ratings per response recording). Table 4 presents the correlation coefficients.

Human–human and human–machine scatter plots for passages are presented in Figures 1 and 2.

The ALSA and NAAL participants are presented as two different data series since machine scoring of the two was different: Speech recognition was used for NAAL

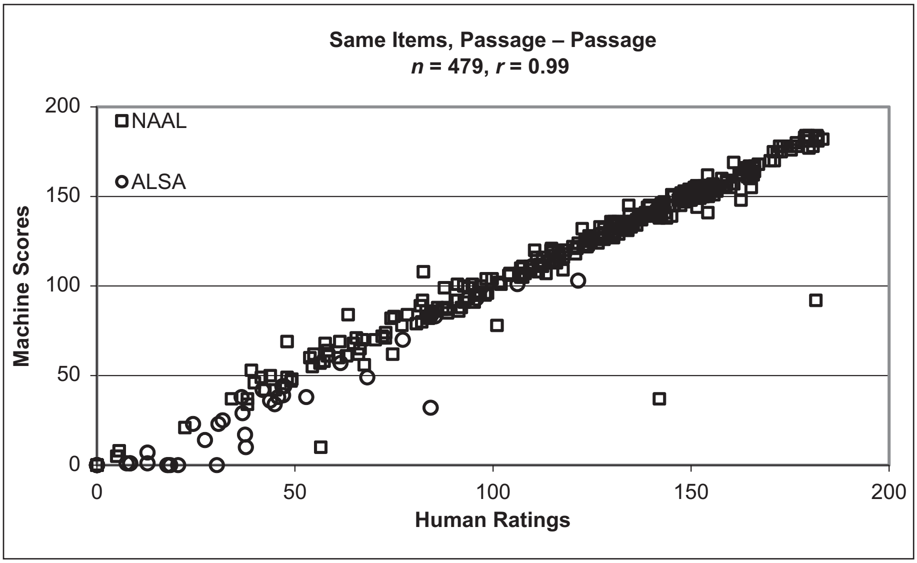

Figure 2. Human versus machine scatter plot for number of words read correctly
Note: NAAL = National Assessment of Adult Literacy; ALSA = Adult Literary Supplemental Assessment.

participants, and transcriptions were used for ALSA participants. The arrangement of the data points in the scatterplots suggests a very strong relation between the two variables in each comparison, with a bit more spread in the sparse data at the low end of Figure 2.

Using generalizability theory (Brennan, 2001), average human ratings and machine scores on passages were analyzed. The variance components in this analysis provided estimates of the relative effects of three main facets: (a) participant ability, (b) the different tasks (different passages), and (c) the raters (human vs. machine). Estimates were also calculated for interactions between the facets: (d) participant-by-task interaction, (e) participant-by-rater interaction, (g) task-by-rater interaction, and (h) participant-by-task-by-rater interaction, which is confounded with unsystematic error not accounted for by the other variance components (see Table 5). Any of the facets that contributed more than 5% of the variance to the model were deemed to have had a significant impact on reading scores. The variance associated with the participant facet shows the difference in scores from participant to participant averaged across raters and tasks. The participant facet accounted for 53% of the variance, which indicates that there were significant differences in abilities across participants as expected. The task facet shows the variation in difficulty of passages averaged across participants and raters. Significant variance was attributed to the task facet (11.5%), suggesting that some passages were more difficult than others. The rater facet contributed a nonsignificant amount of variance to the model (0%). Averaged across participants and tasks, the human raters and machine scores were highly comparable and did not account for any additional variance.

Table 5. Variance Components of a Model of Average Human Ratings and Machine Scores

	Model variance	Standard error	Percentage of V
Participant	1828.32	162.45	53
Task	403.30	331.38	11.5
Rater	0.03	0.06	0
Participant × Task	1211.31	78.71	35
Participant × Rater	7.23	1.11	0
Task × Rater	0.02	0.04	0
PRT, error	16.09	1.04	0.5

Table 6. Generalizability Coefficients From a Model of Average Human Ratings and Machine Scores (Baseline Values From a Model With Individual Human Ratings)

	1 task	2 tasks	3 tasks	4 tasks
1 rater	0.60 (0.60)	0.75 (0.75)	0.81 (0.82)	0.85 (0.86)
2 raters	0.60 (0.60)	0.75 (0.75)	0.82 (0.82)	0.86 (0.86)
3 raters	0.60 (0.60)	0.75 (0.75)	0.82 (0.82)	0.86 (0.86)
4 raters	0.60 (0.60)	0.75 (0.75)	0.82 (0.82)	0.86 (0.86)

The next facet was the interaction between participants and tasks, averaged across raters. The participant-by-task interaction contributed 35% to the model's total variance, which indicates that one passage may have been difficult for Participant A and easy for Participant B, but that another passage may have been easy for Participant A and hard for Participant B. In other words, there was a shuffling in rank ordering across passages. The other two interactions involving raters (participant-by-rater interaction and task-by-rater interaction) did not contribute any variance to the model. The final interaction (participant-by-task-by-rater) refers to the way participants changed across a combination of tasks and raters and includes undifferentiated error. The final interaction only contributed 0.5% to the total model's variance, indicating that there is very little error in the model.

The G-study statistics provided the basis for a D-study (Brennan, 2001), which produced estimates of the population values for different sources of variation in the model. Here, the variance components were used to calculate generalizability coefficients (G-coefficients). By varying the *n* sizes of the number of tasks and raters, the G-coefficients for four tasks and four raters were estimated. Table 6 shows a matrix of G-coefficients for one through four tasks and one through four raters. Table 6 also includes baseline values in which only human ratings were used in the calculations.

Since ratings from human raters and scores from VersaReader were in almost complete agreement, adding more raters did not significantly increase reliability. In contrast, having participants perform more tasks did increase reliability. Overall, the correlation coefficients of the human ratings versus machine scores were almost identical to the baseline of human versus human comparisons.

Table 7. Human Ratings Versus Machine Scores for Passage Readings

		Number of words read correctly		
Participant group	Count	Mean	Standard deviation	Standard error
SP				
Human ratings	190	116	42	3.0
Machine scores	190	116	44	3.2
AA				
Human ratings	189	122	44	3.2
Machine scores	189	121	45	3.3
OE				
Human ratings	193	132	41	2.9
Machine scores	193	130	43	3.1

Note: SP = Native Spanish speakers; AA = Native English-speaking African Americans; OE = Other English speakers.

Machine Bias. The final question to be answered by the validation experiment was whether or not a bias with regard to participants' linguistic/ethnic groups was apparent in the machine scores. The means and standard deviations of machine scores and average human ratings for passage readings from the three participant groups are presented in Table 7. Responses in which the participant read the wrong passage (7 responses) and in which there was no audio (19 responses) were removed from the analysis.

A two-way ANOVA was used to analyze the data. For the two-way ANOVAs, there was one between factor, Participant Group (SP, AA, and OE), and one within factor, Rater Type (human, machine). Alpha for all statistical tests was .05.

For passages, the ANOVAs revealed a statistically significant main effect of participant group, $F(2, 569) = 7.1$, $\mathrm{MS}_{\mathrm{Participant}} = 28{,}695$, $p < .01$. This is consistent with previous findings with human ratings in which the participant groups were different with regard to oral reading ability.

There was no main effect of rater type, $F(1, 569) = 1.8$, $\mathrm{MS}_{\mathrm{Rater}} = 68.6$, $p = .18$. In other words, the machine-generated scores were not consistently higher or lower than scores from the human raters.

Moreover, there was no statistically significant interaction for either of the measures, $F(2, 569) = 2.5$, $\mathrm{MS}_{\mathrm{Participant\ Rater}} = 95$, $p = .08$. The absence of an interaction indicates that no scoring bias was detected in the machine scores with regard to participants' linguistic/ethnic groups. Although the interaction was not statistically significant, there was a trend in the direction of scoring other English speakers slightly more severely than the human raters, while matching the human raters' scores for native Spanish speakers. However, even if the result were statistically significant, the magnitude of the effect size was close to zero ($\omega^2 = .0006$).

The patterns for word lists were similar. ANOVAs for the word list data revealed a statistically significant main effect of participant group, $F(2, 348) = 7.5$, $\mathrm{MS}_{\mathrm{Participant}} = 1{,}187$, $p < .01$. There was no main effect of rater type, $F(1, 348) = 0.42$, $\mathrm{MS}_{\mathrm{Rater}} = 0.56$,

p = .52. There was also no statistically significant interaction, F(2, 348) = 0.1, $\text{MS}_{\text{Participant Rater}}$ = 0.1, p = .90.

Discussion

The correlations of machine scores and average human ratings were almost identical to the correlations between one human's ratings and the remaining human ratings for a given response. In addition, raters contributed 0% of the variance for a model of human ratings and machine scores, indicating that the scores between humans and VersaReader were almost identical. Generalizability coefficients corroborated this finding. For a model with machine scores compared to a model with human ratings reliability coefficients were almost exactly the same and showed that the addition of more raters did not improve reliability. These findings lend support for the machine's ability to score oral reading responses as accurately as an expert human rater.

Statistical analyses showed that the participant groups performed differently on the task. This result was observed previously from the human ratings; namely, the African American group's scores were higher than the native Spanish speaking group's scores, and the other English speaking group's scores were higher than the African American group's scores.

Consistent with the correlation results is the fact that the ANOVA showed that machine scores were not statistically different from the human ratings. The ANOVA also detected no statistically significant bias with regard to participants' linguistic/ethnic group. Although the result of the interaction was not statistically significant, there was a trend in the data for passages. The average machine scores and the human ratings for the native Spanish speaking group were identical, but the average scores for other English speakers were two points lower. The trend is in the direction of scoring Other English speakers slightly more severely than the human raters, while matching the human raters' scores for native Spanish speakers. This direction of severity may be more desirable than artificially penalizing Spanish speakers who are thought to be most often subjected to negative bias. However, the two point difference was unsubstantial when considering that the magnitude of the effect size approached zero.

Conclusion

The four questions addressed in the validation experiment were (a) how reliable were the ratings from human raters, (b) did the human raters introduce bias with regard to participants' linguistic/ethnic groups, (c) how comparable were the machine scores to human ratings, and (d) did the machine scoring introduced bias?

With regard to reliability, the human ratings were extremely consistent when considering both intrarater reliability and intragroup reliability coefficients. Averages for both reliability computations were between 0.99 and 1.00.

Analyses of the human ratings showed no difference in the human ratings across the rater groups and showed no scoring bias.

When considering comparisons between the machine scores and human ratings, the findings indicated that the machine scores were almost indistinguishable from human ratings both for same item correlations and G-coefficients.

Finally, statistical analyses of the machine scores and the human ratings showed no statistical difference between the machine scores and the human ratings and no statistically significant bias in the machine scores.

The results suggest that measuring a participant's reading ability using VersaReader is almost identical to using human ratings. The use of nonnative acoustic models and the detection of reading errors and disfluencies from language models trained on FAN data contribute to the close alignment between VersaReader scores and human ratings.

The findings of the validation experiment have important implications for large-scale assessments. With VersaReader technology, hundreds of thousands of oral reading responses can be scored accurately and efficiently. Automated scoring avoids concerns of inconsistent human rating and biased scoring. The technology eliminates the need for continuous training or calibration of human raters and infrastructure to ensure sufficient intra- and interrater reliability of ratings over time and location. In addition to the advantages of consistent scoring and reduced costs, VersaReader technology has the added benefit of extracting detailed information about the participant's prosodic patterns during oral reading such as the number and length of pauses between words. This type of prosodic information is easy to measure when using speech processing technology but is nearly impossible for human raters to capture in traditional reading assessments. With these benefits, measurement tools like VersaReader might overcome some logistical challenges and enable the inclusion of fluency measures in large-scale assessments.

Even though there is great promise surrounding the use of speech processing technologies for assessing reading ability, the largest obstacle remaining is skepticism that a machine can do as good a job, if not better, than a human rater. The hope is that with more empirical evidence supporting the validity of machine-generated scores, resistance to automation will ease and the benefits to students, teachers, and administrators of using technology for scoring cognitive measures will be embraced not just by the research community but by society at large.

Declaration of Conflicting Interests

All the authors at one point were, or currently are, employees of the company that developed the VersaReader technology described in this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by U.S. Department of Education Contract ED-01-CO-0018 to Westat, under a subcontract to Ordinate Corporation.

References

Adams, M. J. (2006). The promise of automatic speech recognition for fostering literacy growth in children and adults. In M. C. McKenna, L. D. Labbo, R. D. Kieffer, & D. Reinking (Eds.), *International handbook of literacy and technology* (Vol. 2, pp. 109-128). Mahwah, NJ: Erlbaum.

Armbruster, B. B., Lehr, F., & Osborn, J. (2001). *Put reading first: The research building blocks for teaching children to read.* Ann Arbor: University of Michigan, School of Education, Center for the Improvement of Early Reading Achievement.

Baldi, S. (Ed.). (2009). *Technical report and data file user's manual for the National Assessment of Adult Literacy* (NCES 2009-476). Washington, DC: Government Printing Office. Retrieved from http://nces.ed.gov/pubs2009/2009476_1.pdf

Black, M., Tepperman, J., Lee, S., & Narayanan, S. (2008). *Estimation of children's reading ability by fusion of automatic pronunciation verification and fluency detection*. Paper presented at the proceedings of InterSpeech, ICSLP, Brisbane, Queensland, Australia.

Brennan, R. L. (2001). *Generalizability theory.* New York, NY: Springer-Verlag.

Cheng, J., & Townshend, B. (2009). *A rule-based language model for reading recognition*. Paper presented at the proceedings of SLaTE (Speech and Language Technology in Education) 2009, Wroxall Abbey Estate, England.

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* (NCES 2006-469). Washington, DC: Government Printing Office. Retrieved from http://nces.ed.gov/nationsreportcard/pubs/studies/2006469.asp

Leslie, L., & Caldwell, J. (2001). *Qualitative reading inventory-3*. New York, NY: Addison Wesley Longman.

Mostow, J., Aist, G., Huang, C., Junker, B., Kennedy, R., Lan, H., . . . Wierman, A. (2008). 4-Month evaluation of a learner-controlled reading tutor that listens. In V. M. Holland & F. P. Fisher (Eds.), *The path of speech technologies in computer assisted language learning: From research toward practice* (pp. 201-219). New York, NY: Routledge.

Mostow, J., Roth, S. F., Hauptmann, A. G., & Kane, M. (1994, March). *A prototype reading coach that listens*. Paper presented at the proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), Seattle, WA.