Automatic Evaluation of Reading Accuracy: Assessing Machine Scores

Jennifer Balogh, Jared Bernstein, Jian Cheng, Brent Townshend

Ordinate Corporation
Menlo Park, California, USA
jian at ordinate.com

Abstract
Ordinate developed an automatic assessment of oral reading fluency that was administered to a large sample of American adults. Because fluent reading entails accuracy, the machine evaluations of oral reading accuracy were assessed. This paper reviews the methods and results of a study to assess accuracy and bias within a large-scale automatic assessment of oral reading fluency. An experiment compared machine scores with human ratings to measure accuracy and detect any bias for linguistic/ethnic groups. The individual data products of the machine scores are described and the validation experiment is presented. The machine scores were substantially identical to the human ratings.

Index Terms: reading, accuracy, speech recognition, bias

1. Introduction
The U.S. Government’s National Center for Education Statistics commissioned an assessment as part of the 2003 NAAL (National Assessment of Adult Literacy) called the Fluency Addition to NAAL (FAN). For the FAN, each respondent read aloud from lists and passages of text. The oral reading responses were digitally recorded and subsequently analyzed for measures of accuracy and fluency.

Even though oral reading fluency is an important component of general reading ability, it is often not included in large-scale assessments due to the time and expense needed to collect and analyze the data. For the 2003 NAAL project, 181,420 FAN response recordings were collected. To expedite the scoring of these responses and to extract additional information from the responses that human raters cannot provide, technology from Ordinate Corporation was used to automate and augment the analysis of the oral readings.

This paper presents evidence that Ordinate’s automatic scoring of reading accuracy is reliable and valid based on the results of a validation experiment that compared machine scores with human ratings and determined if a bias existed in the way humans and/or the machine assessed respondents from specific linguistic/ethnic groups.

2. General methods
Although the characterization of an individual’s level of literacy can encompass both accuracy and fluency, the focus of the analysis for this report is accuracy. Ordinate was contracted to produce a Span Summary that provided basic measures of oral reading fluency for each of the 18,142 NAAL respondents.

The Span Summary included six primary data products for each response recording.
1. Span summary
   - index of first word attempted
   - index of last word attempted
   - number of words read correctly
   - narrow time (in minutes)
2. Articulation rate
3. Number of short pauses
4. Number of long pauses
5. Number of words completely deleted
6. Number of false starts

However, in order to establish the validity of the reading fluency measures reported in the Span Summary, Ordinate needed to establish the basic accuracy of its scoring method to determine if an automatic system can fairly estimate the number of words that a person can read aloud per time, independent of that person’s dialect or first language?

2.1. Full Data Set
The sample of respondents surveyed in the FAN reflected the demographic profile of the United States. Each of the 18,142 adult respondents read aloud a set of lists and text passages as described in Table 1.

Table 1. FAN material for subject tasks.

<table>
<thead>
<tr>
<th>Tasks (texts)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Digit List</td>
<td>35 one-digit numbers in quasi-random order.</td>
</tr>
<tr>
<td>1 Letter List</td>
<td>35 letters also in quasi-random order.</td>
</tr>
<tr>
<td>3 Word Lists</td>
<td>Each list had 42 words presented in three columns. Lists &amp; words within lists were presented in increasing order of presumed difficulty.</td>
</tr>
<tr>
<td>3 Pseudo-Word Lists</td>
<td>Each pseudo-word list comprised 30 pseudo-words, presented in three columns. Lists &amp; pseudo-words within lists were presented in increasing order of presumed difficulty. (e.g. list 1: dag, list 2: phraw, list 3: vortastic.)</td>
</tr>
<tr>
<td>2 Passages (of 8)</td>
<td>Passages were relatively simple expository or narrative texts of about 150-200 words in length. Comprehension questions were not scored.</td>
</tr>
</tbody>
</table>
The operational definition of ‘accurate oral reading’ used in this project was rendering text into a spoken form that corresponds to a likely rendition of the text by highly literate members of the reader’s community, and from which the text could, in principle, be reproduced verbatim by a literate listener. This definition implies that many different renditions of the same text may be completely accurate, as long as they are intelligible to appropriate listeners.

2.2. Accuracy Validation Experiment

Ordinate’s validation of the oral reading accuracy for the FAN assessment addressed four specific questions:
1. How reliable are human ratings as to both intra-rater and inter-rater agreement?
2. Are machine-generated scores comparable to human judgments on equivalent measures using standard practices for such measures?
3. Is there a significant interaction between human raters’ linguistic/ethnic classification and how raters score the performances of respondents in the different groups?
4. Do the machine-generated measures exhibit a bias with respect to respondents’ linguistic/ethnic classification?

The framework for the human assessment of accuracy was based on the Qualitative Reading Inventory-3 (Leslie & Caldwell, 2001), a traditional oral reading assessment method. After reviewing several widely-used assessments of oral reading accuracy (e.g., Good, Kaminski & Dill, 2002; Swanson et al. 2003; Texas Primary Reading Inventory, 2002), the QRI-3 was chosen as being representative of common practices in human-judged reading assessment. Ordinate’s goal was to use an example of current good practice to evaluate a new machinescoring approach. The QRI-3 is a widely used inventory with well described procedures. The QRI-3 classification of errors includes substitutions, insertions, omissions, reversals, and self-corrections.

One of the goals of the validation experiment was to determine whether the human raters or the machine introduced any detectable bias for or against respondents of various linguistic groups. For this purpose, three linguistic/ethnic groups were identified:
- African American English native speakers (AA)
- Native Spanish speakers (SP)
- Other English native speakers (OE)

2.2.1. Method

Ten human raters active in reading education participated in the experiment. Raters were assigned to one of the three groups (AA, SP, or OE) 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, four were classified as AA, three as SP, and three as OE.

2.2.2. Materials

Response recordings from an experimental sample of 480 respondents from the FAN data collection were analyzed in the validation experiment. Of the 480 respondents, 160 were from each of the three linguistic/ethnic groups. The 480 respondents were selected from their respective demographic groups in a constrained random draw. First, the entire NAAL/ALSA respondent sample was stratified into 10 bins based on the percentile score on a different series of literacy tasks. Sixteen respondents from each linguistic/ethnic group (AA, SP, and OE) were sampled from each of the 10 bins for the validation experiment so that readers with similar ranges of abilities from each of the three groups were sampled for the experiment.

Two response recordings from each of the 480 respondents were analyzed. For 50% of the respondents (240), two passages were analyzed in the study, for 25% (120 respondents) two word lists were analyzed, and for the remaining 25% (120 respondents) 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 FAN data collection. Participants were randomly assigned to each of the three groups: passage-passage, word list-word list, and passage-word list. Table 4 presents the types of responses analyzed in the validation experiment by respondent linguistic/ethnic groups.

Table 2. Response recordings analyzed; p stands for passages, w for word lists.

<table>
<thead>
<tr>
<th>Group</th>
<th>Respondents</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-p</td>
<td>240</td>
<td>480</td>
</tr>
<tr>
<td>w-w</td>
<td>120</td>
<td>240</td>
</tr>
<tr>
<td>p-w</td>
<td>120</td>
<td>240</td>
</tr>
<tr>
<td>Total</td>
<td>480</td>
<td>960</td>
</tr>
</tbody>
</table>

Thus, there are 600 passage recordings and 360 word list recordings in the experimental data set.

2.3. Results

2.3.1. Comparison of Machine Scores & Human Ratings

To assess if machine scores are comparable to human ratings, scatterplots were generated and correlations were calculated for each FAN item analyzed in the validation experiment. The measure used was number of words read correctly. All results are shown in Table 3, and example scatter-plots for word passage 61-1 and list 64-4 appear in Figure 1.

The high correlations (0.96 to 1.00) from this initial set of comparisons suggest that the machine scores and human ratings are very similar, both for the passages and for the word lists.

Each respondent’s data were then assigned to one of three groups depending on whether the respondent’s two response recordings analyzed in the experiment were both passages, both word lists, or one of each. Specifically, the three groups were passage-passage, word list-word list, and passage-word list.

For each group, correlations were calculated for the same items and also cross items (meaning the two different response recordings from the same respondent). Cross items were of interest because, apart from any measurement error, respondents themselves introduce variability across items of the same type and across item types (passages versus word lists). The primary trait of interest is the person’s reading ability as it manifests itself across particular tasks. Ideally, machine scores will be as reliable as the human ratings in assessing the respondent’s general reading ability across items.
Table 3. Correlations of human ratings and machine scores of number of words read correctly for each task.

<table>
<thead>
<tr>
<th>FAN task</th>
<th>Count</th>
<th>Human-Machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>passage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>61-1</td>
<td>75</td>
<td>0.96</td>
</tr>
<tr>
<td>61-4</td>
<td>75</td>
<td>0.98</td>
</tr>
<tr>
<td>61-6</td>
<td>74</td>
<td>0.99</td>
</tr>
<tr>
<td>61-8</td>
<td>75</td>
<td>0.97</td>
</tr>
<tr>
<td>61-9</td>
<td>75</td>
<td>1.00</td>
</tr>
<tr>
<td>61-11</td>
<td>74</td>
<td>1.00</td>
</tr>
<tr>
<td>61-12</td>
<td>75</td>
<td>0.99</td>
</tr>
<tr>
<td>61-16</td>
<td>75</td>
<td>1.00</td>
</tr>
<tr>
<td>word list</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64-2</td>
<td>120</td>
<td>0.98</td>
</tr>
<tr>
<td>64-3</td>
<td>120</td>
<td>0.99</td>
</tr>
<tr>
<td>64-4</td>
<td>120</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Figure 1 Human Machine score relation

Each response recording was rated by three to seven human raters and by the machine. For all human-human correlations, a single rater’s score was compared to the average of the scores from the other raters who rated that response recording. For the human-machine correlations, an average of all available human rater scores was used (from three to seven scores per response recording).

Among several measures of machine accuracy, we report here only on the number of words read correctly. Two other related measures (an accuracy ratio, and the accurate reading rate or WCPM) were analyzed in parallel, with parallel results. The different permutations of rater type (human vs. human, machine vs. machine, and human vs. machine), item type (same items, cross items), and task type (passage-passage, word list-word list, and passage-word list) resulted in 18 different correlations and scatter plots per measure.

The data points in the same-item scatterplots for number of words read correctly are all close to the diagonal. Consistent with this are the very high correlations for each comparison. The data points in the cross-item scatter-plots show much more dispersion. This is true regardless of the type of rater, human or machine. Similar patterns hold true for word list-word list and passage-word list comparisons. These results suggest that measuring the respondent’s reading ability using human ratings is almost identical to using machine scores.
2.3.2. Human Rater Bias

The third question addressed by the validation experiment was whether or not the human raters introduced a bias with regard to the respondents’ linguistic/ethnic classification.

To determine whether there were significant main effects of respondent group, rater group, and/or an interaction between the two, a two-way analysis of variance (ANOVA) was used. The two factors were Rater Group (AA, SP, and OE) x Respondent Group (AA, SP, and OE). The ANOVAs revealed a significant main effect of respondent group for number of words read correctly, $F_{(2,1701)} = 20.9$, $p < 0.01$. The main effect merely demonstrates that the oral reading performance scores of the three respondent groups differed. The main effect of rater group was not significant and there was also no significant interaction. Finding no interaction indicates that no scoring bias was detected in the human ratings.

2.3.3. Machine Bias

The final question was whether or not a bias for linguistic/ethnic groups was apparent in the machine scores. Means and standard deviations of machine scores and average human ratings for passages from the three groups are presented in Table 4.

Table 2. Human vs. machine scores for passages.

<table>
<thead>
<tr>
<th>Group</th>
<th>Score</th>
<th>Count</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Std. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>Human</td>
<td>189</td>
<td>122</td>
<td>44</td>
<td>3.2</td>
</tr>
<tr>
<td>AA</td>
<td>Machine</td>
<td>189</td>
<td>121</td>
<td>45</td>
<td>3.3</td>
</tr>
<tr>
<td>SP</td>
<td>Human</td>
<td>190</td>
<td>116</td>
<td>42</td>
<td>3.0</td>
</tr>
<tr>
<td>SP</td>
<td>Machine</td>
<td>190</td>
<td>116</td>
<td>44</td>
<td>3.2</td>
</tr>
<tr>
<td>OE</td>
<td>Human</td>
<td>193</td>
<td>132</td>
<td>41</td>
<td>2.9</td>
</tr>
<tr>
<td>OE</td>
<td>Machine</td>
<td>193</td>
<td>130</td>
<td>43</td>
<td>3.1</td>
</tr>
</tbody>
</table>

The means for human ratings and machine scores were very similar. To detect machine bias for linguistic/ethnic groups, a two-way ANOVA had one within factor: Rater Type (human, machine), and one between factor: Respondent Group (AA, SP, and OE). The ANOVAs revealed a significant main effect of respondent group, which is consistent with human ratings, but no main effect of rater type; i.e., the machine scores were not consistently higher or lower than human scores. Moreover, there was no significant interaction for any of the measures.

Absence of an interaction indicates that no scoring bias was detected in the machine scores for respondents’ linguistic/ethnic groups. The patterns for word lists were similar.

Overall, the analyses performed to detect machine bias revealed that there was a significant difference in oral reading ability among the three respondent groups as observed previously with human ratings; the machine scores were not significantly higher or lower than the human ratings; and no machine bias with regard to respondents’ linguistic group was detected.

3. Conclusions

The four questions addressed in the validation experiment were (1) how reliable were the human raters, (2) how comparable were the machine scores to human ratings, (3) do human raters introduce a bias based on linguistic/ethnic groups, and (4) does the machine introduced a bias.

The human ratings were very reliable – both intra-rater (range: 0.92 – 1.00) and inter-rater (range: 0.93 – 1.00).

Findings indicated that machine scores were as highly correlated with the human ratings as a single human rater’s scores were with other human raters. No human-machine correlation was significantly lower than the human-human correlation baselines.

Analyses of the human ratings across the three respondent groups (AA, SP, and OE) revealed a significant difference in performance levels of the respondents; namely, OE respondents performed better than AA respondents, who in turn performed better than SP respondents. Results showed no difference in the human ratings across rater groups and no observed scoring bias.

Finally, statistical analyses of the machine scores and human ratings across the three respondent groups (AA, SP, and OE) showed that performance levels of the three respondent groups were significantly different from one another (consistent with the analysis above). The results also revealed no significant difference between the machine scores and human ratings and no bias in the machine scores. These results suggest that machine scoring of oral reading fluency can be as reliable and valid as human scoring for passage- and list-reading tasks.

4. Acknowledgements

Work was performed as a subcontract to Westat, Inc., under a contract with the U.S. Department of Education.

5. References


Texas Education Agency (2001-2002). Texas Primary Reading Inventory (TPRI). Austin, TX: Texas Education Agency.