Predictors of Low Agreement Between Automated Speech Recognition and Human Scores

Joseph F. T. Nese
Josh Kahn
Akihito Kamata

April, 2017
Poster presented at the National Council on Measurement in Education annual meeting

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140203 to the University of Oregon. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

Despite prevalent use and practical application, the current and standard assessment of oral reading fluency (ORF) presents considerable limitations which reduces its validity in estimating growth and monitoring student progress, including: (a) high cost of implementation; (b) tenuous passage equivalence; and (c) bias, large standard error, and tenuous reliability. To address these limitations, the Computerized Oral Reading Evaluation (CORE) system contains an automated scoring algorithm based on a speech recognition engine and a novel latent variable psychometric model. The purpose of this study is to investigate potential student and passage predictors of low agreement between an automated speech recognition (ASR) engine and human scores of words read correctly in student oral reading fluency passages. We fit a cross-classified, variable exposure Poisson model to estimate agreement and found that the majority of variance was found at the student and recording levels, and that student demographic variables explained only a small amount (13%) of the student-level variance.

Conceptual Framework

Assessing oral reading fluency (ORF) is critical because it functions as an indicator of comprehension and overall reading achievement (e.g., Deno, 1985; Hosp & Fuchs, 2005; Marston, 1989). Research indicates that reading fluency should be regularly assessed in the classroom so an instructional response can be made when a difficulty is identified (e.g., Snow, Burns, & Griffin, 1998). ORF curriculum-based measurement (CBM) is used to identify students at-risk for poor learning outcomes through screening assessments, and to monitor student progress to help guide and inform instructional decision-making (e.g., Fuchs, Fuchs, Hosp, & Jenkins, 2001; Speece, Case, & Molloy, 2003).

Despite prevalent use and practical application, the current and standard assessment of ORF presents considerable limitations which reduces its validity in estimating growth and monitoring student progress, including: (a) high cost of implementation; (b) tenuous passage equivalence; and (c) bias, large standard error, and tenuous reliability.

To address these limitations, the Computerized Oral Reading Evaluation (CORE) system contains an automated scoring algorithm based on a speech recognition engine and a novel latent variable psychometric model. Recent research on this system has shown that (a) mean error rates (proportion of words that were scored as incorrect) for a passage were highest for ASR (Table 1), (b) the agreement rate (kappa; Cohen, 1960) between ASR and human scores was about .88, on average, for both students and passages, but the *SD* was quite different (Table 2; about .15 for students and .03 for passages; Nese, Alonzo, Kamata, 2016; Nese, Kamata, Alonzo, 2015). The purpose of this study is to build upon prior research and investigate potential predictors of low agreement between ASR and human word scores.

The purpose of this study is to investigate potential student and passage predictors of low agreement between an automated speech recognition (ASR) engine and human scores of words read correctly in student oral reading fluency passages. We fit a multi-level, cross-classified IRT model to model a latent estimate of agreement.

Research Questions

This study investigates potential student and passage predictors of low agreement between an automated speech recognition (ASR) engine and human scores of words read correctly in student oral reading fluency (ORF) passages. Our research questions are:

(1) How is the variance in latent agreement estimates partitioned at the student and passage levels?

(2) What student and passage variables predict latent agreement estimates?

Methods

Sample. The sample includes 560 students in Grades 2, 3, and 4 across two school districts in Oregon. See Table 3 for sample descriptive statistics.

Measures. The traditional ORF measures were taken from the easyCBM online screening and progress monitoring assessment system (Alonzo, Tindal, Ulmer, & Glasgow, 2006). Each passage was created to be consistent in length (≈250 words) and the readability of each form was verified to fit appropriate grade-level, initially using the Flesch-Kincaid index (e.g., Alonzo & Tindal, 2008), with later empirical support through applications in the field.

The CORE passages are original works of fiction, ±5 words of the target word length (short ≈ 25, medium ≈ 50, long ≈ 85). Passages were written with grade-appropriate vocabulary and word frequency so that an average of several well-respected readability scores was estimated to be at grade-level.

The word accuracy (correct or incorrect) of all passages was scored by trained human assessors via audio recordings (human), and an automated speech recognition engine (ASR). All students were administered the passages via computer: one traditional ORF passage, and 15-18 CORE passages (2-3 long; 3-5 medium; and 8-10 short).

Analysis. To explore the factors that may contribute to poor agreement between machine and human scores, we fit a cross-classified variable exposure Poisson model

$$\log(\lambda) = \beta x + \log(\omega)$$

Where λ is disagreement between ASR and human scores (0 = both scored the word read as correctly or incorrectly; 1 = one scored word as correct, the other as incorrect), and ω is the exposure variable (total number of words per recorded audio), with random effects for the student and passage, and fixed effects for student gender, disability status, and English Learner (EL) status, and recording duration. We used the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) in the R programming language (R Core Team, 2016) to conduct the analyses.

The baseline model (`m0`) was specified as follows:

```
m0 <- glmer(disagree_sum ~ 1 + offset(log(total_words)) +
(1|recording) + (1|student) + (1|passage), family = poisson)
```

The comparison model (m1) was specified as follows:

```
m1 <- glmer(disagree_sum ~ 1 + offset(log(total_words)) + gender
+ disability + el + recording_duration + (1|recording) +
(1|student) + (1|passage), family = poisson)
```

Results

Results showed that `m1` explained approximately 10% of the `m0` variance, with: no variance explained at the recording-level; gender, disability status, and EL status explaining 13%

of variance at the student-level; and recording duration explaining 53% of variance at the passage-level (see Table 4). Note that additional passage covariates (e.g., Flesch-Kincaid, average word length) accounted for an additional 4% of variance at the passage-level.

See Table 5 for the fixed effect model results. The rate of disagreement for the intercept (female, non-disability, non-ELL, Grade 3, average recording duration) was 0.06. All else constant, the rate of disagreement was 1.86 times higher for students with disabilities than students without disabilities. All else constant, the rate of disagreement for Grades 3 and 4 were about half the rate for Grade 2. For a standard deviation increase in the recording duration (18.5 seconds), the rate of disagreement rate increased by about 10%.

Conclusion

In response to our first research question, there was only a small proportion of variance at the passage level; the majority of variance was found at the student and recording levels. In response to our second research question student demographic variables explained a moderate amount of the student-level variance (13%), and we were unable to explain any of the variance associated at the recording level (note that we did have a human rating of "audio quality" for about 30% of the recordings, but this variable did not reduce variance at any level in a meaningful way). The results of this study have the potential to begin to understand how the ASR scores readings of English learners or students with disabilities, to inform the refinement of the CORE system by identifying predictors that may indicate (*a priori*) an unreliable ASR score, and to identify text properties that degrade ASR scoring so that future models can be trained on these features.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67*, 1-48. doi:10.18637/jss.v067.i01

Cohen, J. A. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 18,* 19–32.

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.

Hosp, M. K., & Fuchs, L. S. (2005). Using CBM as an indicator of decoding, word reading, and comprehension: Do the relations change with grade? *School Psychology Review, 34,* 9–26.

Nese, J. F. T., Alonzo, J., & Kamata, A. (2016, April). *Comparing passage lengths and human vs. speech recognition scoring or oral reading fluency.* Paper presented at the annual meeting of the American Educational Research Association (AERA), Washington, DC.

Nese, J. F. T., Kamata, A., & Alonzo, J. (2015, July). *Exploring the evidence of speech recognition and shorter passage length in Computerized Oral Reading Fluency (CORE).* In K. Cummings (Chair), *Assessment fidelity in reading research: Effects of examiner, reading passage, and scoring methods.* Symposium conducted at the Society for the Scientific Study of Reading (SSSR), Hawaii.

R Core Team (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Speece, D. L., Case, L. P., & Molloy, D. E. (2003). Responsiveness to general education instruction as the first gate to learning disabilities identification. *Learning Disabilities Research & Practice, 18,* 147-156.

Table 1

Comparisons of Error Rate Means and (SD) across Human and ASR, and Three CORE Passage Lengths

		Short 25 words	Medium 50 words	Long 85 words
Grade 2 (*n* = 127)	Human	.09 (.10)	.05 (.07)	.05 (.07)
	ASR	.11 (.13)	.08 (.10)	.09 (.10)
Grade 3 (*n* = 158)	Human	.06 (.31)	.05 (.07)	.03 (.05)
	ASR	.07 (.09)	.07 (.09)	.06 (.07)
Grade 4 (*n* = 162)	Human	.07 (.11)	.04 (.08)	.04 (.05)
	ASR	.08 (.12)	.06 (.09)	.05 (.07)

Table 2

Word-level Agreement (Cohen's kappa) Comparisons between Human and ASR at the Student and Passage Levels, Across Grades

	Grade 2		Grade 3		Grade 4	
	Student (n = 127)	Passage (n = 54)	Student (n = 158)	Passage (n = 54)	Student (n = 162)	Passage (n = 52)
Mean	.82	.83	.90	.90	.91	.91
SD	.20	.04	.14	.03	.12	.03
min	.16	.73	.09	.84	.29	.84
max	.99	.90	.99	.96	1.00	.96

Table 3

Sample Descriptive Statistics

	n	%
Grade		
2	147	26
3	146	26
4	267	48
Sex		
Female	278	50
Ethnicity		
Hispanic/Latino	105	19
Race		
American Indian/Alaskan Native	15	3
Asian	10	2
Black	7	1
Multi-Race	27	5
Native Hawaiian/Pacific Islander	1	0
Non-US Native American	3	1
Pacific Islander	2	0
White	495	88
Disability	82	15
English Learners	59	11

Table 4

Random Effect Variance of Models

	m0		m1		
Groups	Variance	% of total m0	Variance	% of total m1	"Explained Variance" (m0−m1)/m0
recordings	0.71	47%	0.73	53%	-2%
students	0.67	44%	0.59	43%	13%
passages	0.14	9%	0.07	5%	53%
Total	1.53		1.38		10%

Table 5

Model Fixed Effects

	Estimate	SE	log(Estimate)
(Intercept)	-3.36*	0.06	0.06
Male	0.03*	0.07	1.03
Disability	0.62*	0.10	1.86
ELL	0.22*	0.11	1.24
Grade 2	0.54*	0.06	1.72
Grade 4	-0.14*	0.07	0.87
Recording duration	0.01*	0.00	1.01

* $p < .001$.