Examining decoding items using engine transcriptions and scoring in early literacy assessment

Zachary Schultz and **Mackenzie Young** and **Debbie Dugdale** and **Sue Lottridge**
Cambium Assessment
{zachary.schultz, mackenzie.young, debbie.dugdale, susan.lottridge}@cambiumassessment.com

Abstract

We attempt to improve a transcription-based early childhood speech scoring approach by implementing allowable variations, which are phonemes that can be swapped out for those in a target word to create adjacent words that would receive a passing score. This approach is based off of how a child acquires the ability to pronounce sounds in the English language and how certain phonemes can be easily confused during transcribing, whether done by a human or a machine. Testing both a set of general allowable variations and a set specifically based on the American southern dialect against human scoring, we find that the general allowable variations improve performance, especially at item-specific levels. The performance when utilizing dialect-specific variations does not change much, although this is likely from the use of a word-based transcriber. This approach can be useful for specific words that may have phonemes easily misheard for their voiced or voiceless counterpart (e.g. "have" and "half") but, overall, a modeling approach for training an AI engine would more likely result in higher performance.

1 Introduction

Automated scoring of text-based items is common in K-12 assessment. Automated scoring of speech-based items is used in English Language learner assessments (e.g., Texas English Proficiency Assessment Program) as well as some early literacy screeners (e.g., Soapbox labs, Amira). In automated scoring of speech-based items, automated speech recognition systems are utilized, sometimes taking a transcriber-based approach. Transcribers, whether human or machine, are prone to mistakes, with many human transcribers requiring professional training in order to achieve accurate and quality transcriptions. Even the most reliable, open-source transcriber model, Whisper, can experience vastly different Word Error Rates (WER) depending on the acoustic environment of the audio and the speaker themselves (Kuhn et al., 2024.)

Standard Whisper-based models are not usually trained on child speech and are therefore more prone to errors when transcribing audio of children speaking (Jain et al., 2023.) When utilizing this transcribing approach for speech scoring, in which speech is transcribed and then a rules-based scoring process is applied, one should take into account linguistic features. In particular, for early literacy assessment or the assessment of young children's speech, understanding how children develop their articulatory skills and how phonemes are connected in their place or manner of articulation can contribute to potentially more robust scoring and results that can more accurately inform about a student's speaking ability.

In bridging the gaps between machine scoring, psychometrics, and linguistics, we explore the ways in which one transcription-based approach could be improved by the use of "allowable variations" in early literacy verbal tasks.

2 Background

Early literacy assessments are becoming a critical piece of K-12 large scale assessment to support evidence-based reading instruction (Brunetti et al., 2025). Most early literacy assessments consider reading fluency as a combination of word recognition and language comprehension (Gough & Turner, 1986; Scarborough, 2001; Duke & Cartwright, 2021). Word recognition can be divided into three broad strands: phonological awareness, sight recognition, and decoding, with the latter being the focus of this study.

Decoding is the process of linking printed letters to spoken sounds and includes recognition of phonology, orthography, and morphology (Clemens et al., 2020). During decoding, readers might sound out and blend individual letters

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 191–196
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

into phonemes or combine larger letter groups to form syllables and recognize whole words (Ehri, 2005). Garcia & Cain's meta-analysis (2014) analyzed decoding assessment characteristics and found that the accurate decoding of real words (vs pseudowords) was more predictive of reading comprehension than other measures.

In one decoding item type, students are shown a word and asked to say it aloud. Scoring involves determining whether the student verbalized the target word accurately and whether and what "variations" are allowed, in order to recognize multiple factors that can influence this determination. For instance, words are typically interpreted by humans (and engines) in the context of other words; without this context, both humans and engines can interpret a word slightly differently with both likely being correct representatives. Additionally, the acoustic and linguistic properties of very young children's speech can impact how both humans and engines interpret pronunciation. Acoustically, children's speech falls in a higher register and can have prosodic characteristics that differ from adults. Linguistically, children have underdeveloped articulatory systems and some may struggle to pronounce more advanced English phonemes. This can lead to children replacing a more difficult sound with an easier one (e.g., /r/ vs. /w/). Finally, dialect or regional pronunciations can impact how words are pronounced.

When tests are administered remotely via computer in a classroom setting, testing conditions can also impact scoring (Oberle & Powers, 2025). Often, tests are administered at the same time within a classroom; there can be substantial background noise and chatter, multiple speakers, as well as variations in how loudly or quickly a student speaks. And, young students' ability to interact with the test can also contribute to the demonstration of their decoding skill. The determination of "allowable variations" thus needs to consider these factors relative to each target word and an acceptable pronunciation.

There are three ways to score these items. First, the student speech can be scored by trained human raters using a rubric. Second, humans or machines can transcribe the student speech, and then apply explicit scoring rules. Third, AI systems can be modeled directly on speech to predict human scoring.

In this study, we aim to explore the second approach while addressing the previously stated factors that can add difficulty to this method.

3 Methods

We use data from seventeen decoding items administered across kindergarten and grade one during a Spring 2024 operational field test in one southern state. Students could earn a score of 1 for a correct pronunciation or a score of 0 if incorrect. A correct score required an exact pronunciation with little to no variations allowed. In these data, responses were scored by trained, human raters and a subset (100 per item) was transcribed by both a human and a Whisper-based model trained on adult speech (Radford et al., 2022.)

In transcribing, neither the humans nor the machine had knowledge of the target word for each item. Once transcribed, a score of 0 or 1 was given depending on if the transcription contained the target word, with a score of 1 indicating that the target word is present. Initially, only the target word can trigger a score of 1. Then, the list of acceptable words expands once allowable variations are added.

To determine acceptable variations, we first look at the literature concerning how young children may differ in their pronunciations of various phonemes and how their articulatory systems develop. For example, children develop the ability to pronounce consonants such as /b, p, m, n, h, w, d/ around two years of age, whereas consonants such as /ɹ, ʒ, ð, θ/ are acquired at an age between five and seven years old (Crowe & McLeod, 2020.) Because of this, students may replace one of these later-stage sounds with one they acquired earlier. We also consider manner and place of articulation, with the assumption that phonemes that are close in one or both traits may be misinterpreted when transcribing. In terms of vowels, those that are close to one another in the physical vowel space can be considered as allowable variations.

Allowable variations are determined and listed using the International Phonetic Alphabet (IPA), a collection of symbols each representing one unique possible sound in human speech. Sounds in English are sometimes composed of multiple letters but represented as one symbol using the IPA. For example, the English sound written as "th" is represented in the IPA either as /ð/ or /θ/ depending on if it is a voiced sound or not. Evidently, the IPA allows one to represent a specific sound with one character and it is therefore useful in both representing allowable variations and in implementing them computationally.

Once the list of allowable variations is determined, we then apply each of the variations to the IPA transcription of a given response. To convert a transcription to its IPA representation, we use the Python package 'eng-to-ipa.'[1] Each phoneme within the transcription is swapped out with each of its acceptable variations until all possible combinations have been created. This leads to a long list of non-existent words, so we cross reference the created list with the Carnegie Mellon University (CMU) Pronouncing Dictionary[2] and only include those produced transcriptions that are valid and present in the dictionary. This both limits the number of targets that we are accepting beyond the given and it accounts for the transcribers acting more in a word-based manner than a phonemic manner.

The list of variations can also be modified to fit specific frameworks, such as the phonetic inventory of a dialect. While some variations in the general list we created may apply to a dialect, focusing on specific features of a given dialect should produce a more specific set of variations. In this study, we create a list of variations using features of the standard American southern dialect. Features such as monophthongization, diphthongization, triphthongization, non-rhoticity, the "pin-pen" vowel merger, and the distinction between words such as "which" and "witch" are all incorporated into the list. In generating the variant targets with this list, we skip the cross reference with the CMU Pronouncing Dictionary to maintain all the features of the dialect even if they lead to non-standard words.

With both a general list of acceptable variations[3] created for each target and a list of variations utilizing the features of the American southern dialect, we then rescore by looping through the variation lists alongside the transcriptions. If any of the variations or the original target word are present in the transcription, a score of 1 is given. If the target word nor any of the variations are present, a score of 0 is given. We then use these scores to calculate comparative statistics in order to gauge changes in item performance.

We use three statistical measures to gauge agreement in this study. Firstly, we calculate exact agreement between two sets of scores. Exact agreement is a percentage of the scores that are the same for a given response for both sets. Secondly, we use Quadratic Weighted Kappa (QWK or Kappa, for short). This is another agreement statistic with a more robust calculation which takes into account the possibility that an agreement occurred by chance. It also penalizes disagreements that are further from one another on an ordinal scale; however, this is irrelevant in this study as there are only two possible labels for the data. Finally, we calculate the p-value, which in this case is the mean score. These measures were then averaged across grade level and averaged overall. We compare agreements between the human raters, the first human rater and both types of transcription, and between the transcriptions. The goal in this study is to be comparable to, or better than, the agreement values between the two human raters, which are outlined in Table 1.

Grade	N		H1H2 Scores		
	Items	Responses	Exact Agr.	K	p-val.
K	3	203	89%	0.74	0.66
1	14	182	93%	0.73	0.81
All	17	186	92%	0.73	0.79

Table 1: Agreement statistics between the two human raters

4 Results

4.1 General Variations

Table 2 provides an overview of the agreement statistics when a score of 1 is strictly given for the target word and no variations are included. One can see that a transcribing method with similar scoring rules to the human raters does not perform as well as humans. The low Kappa values here are primarily due to machine transcription error. This prompted an attempt to improve these results through the use of allowable variations.

Table 3 provides an overview of the agreement statistics when allowable variations are used. These variations are those from a set of general variations. Table 4 shows the agreement statistics when the variations are focused on the American southern dialect.

When using general variations, performance improves. For exact agreements, there is a slight improvement of 1-3%. Similarly, kappa values show slight improvement as well, ranging from an increase of 0.01 to 0.05. Overall, this is not a large difference, but individual items experienced more

[1] https://pypi.org/project/eng-to-ipa/

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[3] See Table 5 for an example of consonant variations utilized in this study.

Grade	**N**		**H1 Score-Human Transcription**			**H1 Score-Engine Transcription**			**Human-Engine Transcriptions**	
	Items	Scores	Exact Agr.	K	p-val.	Exact Agr.	K	p-val.	Exact Agr.	K
K	3	86	78%	0.58	0.45	73%	0.51	0.42	84%	0.65
1	14	89	79%	0.45	0.64	72%	0.36	0.57	78%	0.45
All	17	88	79%	0.47	0.60	72%	0.39	0.54	79%	0.49

Table 2: Agreement statistics without the use of variations

Grade	**N**		**H1 Score-Human Transcription**			**H1 Score-Engine Transcription**			**Human-Engine Transcriptions**	
	Items	Scores	Exact Agr.	K	p-val.	Exact Agr.	K	p-val.	Exact Agr.	K
K	3	86	76%	0.53	0.52	72%	0.47	0.48	83%	0.66
1	14	89	82%	0.48	0.69	76%	0.40	0.62	79%	0.49
All	17	88	81%	0.49	0.66	75%	0.41	0.59	80%	0.52

Table 3: Agreement statistics with the use of general variations

drastic improvement or decay.

For example, an item with the target word of "have" increased from 65% to 71% in exact agreement and from 0.28 to 0.36 in its kappa value. When examining the data, the acceptable variation "half" appears to be responsible for these increases. This suggests that either students have difficulty distinguishing or pronouncing the voiced and voiceless dental fricatives or that these phonemes sound similar on recording and can be hard to distinguish by listeners and transcribers.

Another item with the target word of "what" experienced similar increases in performance. The exact agreement value increased from 82% to 86% and the kappa value increased from 0.55 to 0.57. This improvement mainly came from vowel variations, especially when swapping the phoneme /ə/ with /ɛ/, which produce the words "what" and "wet" respectively. There are also instances in this case where the word "wood" contributes to performance changes, showing, again, that voiced and voiceless consonant pairs can cause problems for students or listeners.

Finally, the item with the target word of "your" gained large improvements. The exact agreement value increased from 74% to 86% and the kappa value increased from 0.31 to 0.35. In this instance, we produced variations that included the removal of the initial sound and this seemed to lead to this performance improvement. This suggests that initial sounds may be missed in the recording or transcribing process or that words with many minimal pairs may have a higher rate of mistakes by students or listeners and transcribers.

4.2 Dialectical Variations

The use of variations which only involve changes associated with the American southern dialect yielded little change. The only item that experienced significant performance improvement was the item with the target word "been." This is most likely due to the "pin-pen merge" feature of the southern dialect in which the vowels in the words "pin" and "pen" are pronounced almost identically.

Most likely, we do not see drastic performance changes using the southern dialect because transcribing was done using a word-based transcriber, which seeks to output a valid English word if possible. While testing was done with a phonetic transcriber, the output was not reliable. The approach with dialectical variations could be very useful in cases where there are reliable phonetic transcriptions and the use of a specific dialect is well documented for the area of testing.

5 Conclusion

Overall, implementing allowable variations can lead to slight overall performance improvements and item-specific improvements ranging from slight to major. When the variations are a general list, the improvements are higher and more widespread across the items. When they are focused in on a dialect, the improvements are minimal. However, this could be due to the use of a word-based transcriber whereas dialects feature varying pronunciations of a word with one acceptable spelling.

With only slight improvement overall, this approach may only be useful in cases where a specific word is being used and it is likely that a phoneme within the word will lead to transcribing errors. This most often seems to be pairs of voiced and voiceless phonemes but can also occur when a word has many minimal pairs.

The other avenue to take when attempting to improve these agreements is to use high-quality,

Grade	N		H1 Score-Human Transcription			H1 Score-Engine Transcription			Human-Engine Transcriptions	
	Items	Scores	Exact Agr.	K	p-val.	Exact Agr.	K	p-val.	Exact Agr.	K
K	3	86	78%	0.58	0.45	73%	0.51	0.42	84%	0.65
1	14	89	80%	0.46	0.65	74%	0.37	0.59	78%	0.49
All	17	88	80%	0.48	0.61	74%	0.40	0.56	79%	0.52

Table 4: Agreement statistics with the use of dialectical variations

human transcriptions and scores and train an AI engine using them. After looking at the results from this study, we believe this would be the recommended route to take if possible.

Limitations

In this study, the sample sizes per item were fairly small and may not have been representative of the student population. We also used a word-based transcriber which has more difficulty in reporting dialectical features of speech.

Acknowledgments

I am deeply grateful to my coauthors for their help in every aspect of conducting this study. Their expertise, time, and dedication were essential to completing this work. I would also like to thank Amy Burkhardt for her guidance and leadership, not just in this work but in all that I do. Finally, a huge thank you to the rest of the ML Team at Cambium Assessment for their leadership and innovation in this field. I am eternally thankful to be a part of the team and the important work we do.

References

Matthew Brunetti, Meredith Langi, and Sarah Quesen. 2025. Are we on the same page? a discussion on the use and misuse of early literacy assessments.

Nathan H Clemens, Kejin Lee, Maria Henri, Leslie E Simmons, Oi-Man Kwok, and Stephanie Al Otaiba. 2020. Growth on sublexical fluency progress monitoring measures in early kindergarten and relations to word reading acquisition. *J. Sch. Psychol.*, 79:43–62.

Kathryn Crowe and Sharynne McLeod. 2020. Children's english consonant acquisition in the united states: A review. *American Journal of Speech-Language Pathology*, 29(4):2155–2169.

Linnea C Ehri. 2005. Learning to read words: Theory, findings, and issues. *Sci. Stud. Read.*, 9(2):167–188.

Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. 2023. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *Preprint*, arXiv:2311.00430.

J Ricardo García and Kate Cain. 2014. Decoding and reading comprehension. *Rev. Educ. Res.*, 84(1):74–111.

Rishabh Jain, Andrei Barcovschi, Mariam Yiwere, Peter Corcoran, and Horia Cucu. 2023. Adaptation of whisper models to child speech recognition. *Preprint*, arXiv:2307.13008.

Korbinian Kuhn, Verena Kersken, Benedikt Reuter, Niklas Egger, and Gottfried Zimmermann. 2023. Measuring the accuracy of automatic speech recognition solutions. *ACM Transactions on Accessible Computing*, 16(4):1–23.

C. Oberle and S. Powers. 2025. Evaluating early literacy tasks: Insights from a mixed-methods study.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

A Appendix

Original Phoneme	Sound in English	Variations
/p/	p	p, b
/b/	b	b, p
/m/	m	m, b
/t/	t	t, d
/d/	d	d, t
/k/	k	k, g, t
/g/	g	g, k, d
/f/	f	f, v
/n/	n	n, d, nd, ŋ
/ŋ/	ng	ŋ, n
/w/	w	w
/j/	y	j, ’ ’
/h/	h	h, ’ ’
/v/	v	v, f
/s/	s	s, z, θ
/z/	z	z, s, ð
/ʃ/	sh	ʃ, s, θ
/ʒ/	si (as in ’vision’)	ʒ, ʃ
/tʃ/	ch	/tʃ/, ʃ, k, t, dʒ
/l/	l	l, w, j
/dʒ/	j	/dʒ/, tʃ, d
/θ/	th (voiceless)	θ, ð, t, f
/ð/	th (voiced)	ð, θ, d, v
/ɹ/	r	r, w, l, ’ ’
/tr/	tr	tr, tʃ, t
/dr/	dr	dr, dʒ, d
/kr/	kr	kr, gr, r, k, g
/gr/	gr	gr, kr, r, g, k
/skr/	skr	skr, sk, kr, s, k, r

Table 5: Example of general consonant variations