Automated Speech Scoring Methods and Results

Cambium Assessment, Inc.

Abstract

This report presents the methods and results of Cambium Assessment, Inc.'s (CAI) automated scoring of responses for a set of screener speaking items administered to English language learners (ELLs) in K–12 assessments. The purpose of using automated scoring is to provide scores to educators more quickly so that ELL students can be identified and placed into English language development services earlier and to reduce costs around scoring. Automated scoring can also help ensure consistent scoring within and across test administrations. Most automated speech engines rely on explicitly defined or algorithmic features to produce both the transcription (i.e., conversion of speech to text) and features used to predict scores (Zechner & Loukina, 2020). The transcription model maps characteristics of the audio signal to speech sounds, such as phonemes, and these are mapped to words. Features can include characteristics of the speech itself, including measures of fluency, pronunciation quality, and prosody; features can also include aspects of language such as grammar errors and patterns of word usage via word-document feature modeling (Metallinou & Cheng, 2014; Zechner et al., 2014). This study extends the current research by illustrating the performance of multi-layer neural networks in both transcription and scoring. The engine architecture in this study relies heavily on the recent transformer approach (Vaswani et al., 2017), as this has produced state-of-the-art results in both automatic speech recognition (Baevski et al., 2020; Hsu et al., 2021; Wang et al., 2021) and text-based scoring (NCES, 2022; Ormerod et al., 2022; Ghosh, Klebanov, & Song, 2020; Matthias & Bhattacharyya, 2020; Riordan et al., 2020). Wav2vec 2.0 (Hsu et al., 2021) and Unispeech (Wang et al., 2021) were used to transcribe the audio. The transcribed text was submitted to two deep learning transformer-based language models, specifically, ELECTRA (Clark et al., 2020) and FNet (Lee-Thorp et al., 2021). Results from both models were then entered into a one-versus-rest logistic regression model to produce the final score. The results suggest that the transformer-based models perform well compared to humans and that the transcriber choice is critical.

Table of Contents

Introduction and Background

This report presents the methods and results of Cambium Assessment, Inc.'s (CAI) automated scoring of responses for a set of screener speaking items administered to English language learners (ELLs) in K–12 assessments. The purpose of using automated scoring is to provide scores more quickly so that ELL students can be identified and placed into English language development services earlier and to reduce costs around scoring. Automated scoring can also help ensure consistent scoring within and across test administrations.

When trained on a highly validated hand-scored sample and evaluated using agreement and score distribution metrics, the use of automated scoring can help ensure that responses are scored against the rubric as intended. Using automated scoring on constructed responses has become increasingly common in K–12 formative, interim, and summative assessment programs (Foltz, Yan, & Rupp, 2020). This use has been supported by the increase in online testing, more sophisticated approaches in hybrid automated/hand-scoring, and improvements in computational linguistics (Lottridge & Godek, 2022; Palermo, 2022; Boyer, 2020; Habermehl, Nagarajan, & Dooley, 2020). In particular, automated speech scoring has been used in ELL assessment (Pearson, 2017) and in the assessment of the Test of English as a Foreign Language (TOEFL®) (Zechner & Loukina, 2020).

CAI's scoring engine, Autoscore^{SP}, relies on deep learning—or multi-layer neural networks—for both its transcription model and its scoring model. To our knowledge, most automated speech engines rely on explicitly defined or algorithmic features to produce both the transcription (i.e., conversion of speech to text) and/or features used to predict scores (Zechner & Loukina, 2020). The transcription model maps characteristics of the audio signal to speech sounds, such as phonemes, and then these are mapped to words and word sequences. The transcription model also produces outputs that can be used as features in scoring. The Educational Testing Service (ETS) SpeechRater generates more than 100 features used for scoring (Zechner et al., 2014; Loukina et al., 2015). Pearson's Versant generates similar types of features, albeit using different approaches (Cheng et al., 2014). Features can include characteristics of the speech itself including measures of fluency (e.g., speaking rate, pausing patterns), pronunciation quality (e.g., confidence in transcription prediction), and prosody (e.g., duration of phoneme pronunciation) (Chen et al., 2018; Metallinou & Cheng, 2014; Cheng et al., 2014). Features can also include aspects of language such as grammar errors and patterns of word usage via word-document feature modeling (Metallinou & Cheng, 2014; Zechner et al., 2014).

In contrast, the Autoscore^{SP} models use multi-layer neural networks to learn features from the data, both for transcription and for scoring, rather than using explicitly generated features. The engine architecture relies heavily on the recent Transformer approach (Vaswani et al., 2017), as this approach has produced state-of-the-art results in both automatic speech recognition (Baevski et al., 2020; Hsu et al., 2021; Wang et al., 2021) and text-based scoring (NCES, 2022; Ormerod et al., 2022; Ghosh, Klebanov, & Song, 2020; Matthias & Bhattacharyya, 2020; Riordan et al., 2020; Rodriguez, Jafari, & Ormerod, 2019; Taghipour & Ng, 2016). The Transformer approach uses a novel architecture and training approach in which a pretrained model 'learns' characteristics of speech or text on a very large dataset, and then is fine-tuned to a particular task, such as scoring. The underlying concept is that the pretrained model, often called a language model, represents language use in context.

In terms of performance, the existing engines model hand-scoring well, producing similar agreement rates to the rates at which humans themselves agree (Metallinou & Cheng, 2014; Zechner et al., 2014). Word error rates on children's speech in similar programs have been published as around 27–39% (Chen et al., 2018; Metallinou & Cheng, 2014; Cheng et al., 2014). On publicly available datasets on adult speech, the neural network-based transcribers have produced error rates ranging from 4% to 25%,

with results varying across different types of recorded audio (Hsu et al., 2021; Bevski et al., 2020; Wang et al., 2021).

In this report, we describe Autoscore^{SP} design, data used to train the models, and the results of the training. Our results indicate solid performance, both relative to hand-scoring and relative to human transcriptions.

Automated Scoring Engine

Autoscore^{SP} scores student speech by first preprocessing the response audio and then sending the processed response to two transcribers, the transcribed output to two score prediction models, and then to an ensemble that combines the two outputs to predict the final score (refer to Figure 1). Prior to transcription, responses are converted from text-encoded audio and then down-sampled into discrete units (16Khz). While other processing can be used in the engine, we found that the transcribers work best when minimal processing is conducted because this matches how the transcribers are trained. Autoscore^{SP} has the capability of using different transcriber-score prediction combinations and more than two combinations of transcribers and scorers.

Transcription Models

The two transcriber architectures (wav2vec 2.0 and Unispeech) that are used in Autoscore^{SP} showed state-of-the-art performance, as measured by word error rate, on the LibriSpeech or Common Voice datasets (Baevski et al., 2020; Wang et al., 2021). The wav2vec 2.0 architecture has also demonstrated quality in transcribing audio when trained in one domain and used in another (Hsu et al., 2021). The transcribers used in the scoring engine came from the Hugging Face library[1] and are instances of the wav2vec 2.0 and Unispeech architectures. The chosen models were used without modification in the speech engine.

We briefly describe the training procedure for the transcriber models as an illustration of their approach and as support for their use. Both transcriber models underwent two training phases. In Phase 1 (pretraining), the models were trained using large datasets of non-transcribed audio. In this phase, speech representations were estimated. These speech representations reflect the contextual relationship between speech segments (i.e., partitioned audio at set timesteps such as 10 or 25 milliseconds) and allow for speech segments to be represented as a lower dimensional space. These representations are similar to those produced for text. Like models used to build text-based contextual representations, the training procedure at this phase masks speech representations for the audio, and then uses the unmasked speech representations and a transformer design (Vaswani et al., 2017) to predict the best representation for the masked speech representation. This approach was coupled with discretized (Boolean) speech representations that were also used to predict the masked representation. These discretized speech representations are intended to represent common speech sounds and have been shown to be associated with English phonemes (Baevski et al., 2021). The final speech representations arise from a procedure that minimizes the differences between the discretized representations and masked audio segments. The Unispeech approach extended the wav2vec 2.0 architecture by

[1] https://huggingface.co/

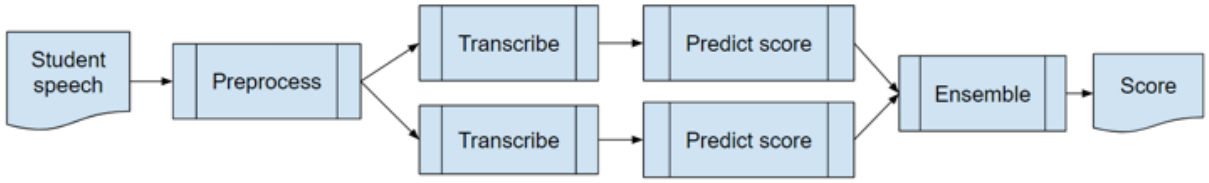

Figure 1. High-level flow of Autoscore^{SP}

using automatically generated phonemic representations and tying them explicitly to the discretized speech representations.

The benefit of the Phase 1 training is to leverage substantial non-transcribed audio to build a general model of speech. At Phase 1, however, the transcriber models have not yet 'seen' human-transcribed text and need to be trained to map the speech representations to this text. This mapping is managed in Phase 2 in a process called fine tuning. In this phase, the speech representations from Phase 1 are used as features that predict characters observed in the transcribed training data. These characters are typically the upper-cased alphabet characters, the apostrophe symbol, a blank symbol, and two system-level tags: PAD, which represents unused text in the audio to accommodate longer and shorter audio responses, and UNK, which represents speech sounds appearing in future audio that are not known in the vocabulary. Because audio segments do not have natural semantic divisions like text (i.e., words, sentences), the mapping needs to consider how to best align the audio file to the character (and subsequent word) predictions. This alignment is managed by the Connectionist Temporal Classification (CTC) loss function, which picks the optimal pattern of character predictions given the audio (and human-transcribed text) by implementing rules around predicted character repetition within defined boundaries (Graves et al., 2006). The training data are at the sentence level and so the CTC loss function is optimizing the match of possible character pattern outputs to the human-transcribed sentence. This optimization considers the order of the characters, and so identifies the best characters in context of the other characters (at the sentence level).

The data used to train the wav2vec 2.0 and Unispeech models for Phase 1 and Phase 2 were from publicly available datasets. Table 1 presents information on the datasets used. Datasets consisted either of read text (typically from books) or conversations by adults. Datasets contained spoken audio by both males and females, and some datasets contained audio with accents (Common Voice). None of the datasets included speech by children. The data are divided typically into sentences (audio and transcribed audio, if applicable) for training.

Table 1. Datasets Used to Train Transcribers, by Transcriber and Training Phase

Phase	Dataset	Description
Wav2vec 2.0[2]		
1	Libri-Light (Kahn et al., 2019)	Open-source audio books read by volunteers[3]
	Common Voice (Ardila et al., 2020)	Crowd-source collected audio data with read-out sentences
	Switchboard (Godfrey & Holliman, 1993)	Telephone speech corpus between two randomly assigned adults on selected topics
	Fisher (Cieri et al., 2004)	Telephone speech corpus between two randomly assigned adults on selected topics
2	Librispeech (Panayotov et al., 2015	Open-source read-out audio book data
Unispeech[4]		
1	Common Voice (Ardila et al., 2020)	Crowd-source collected audio data with read-out sentences
2	TIMIT (Garofolo et al., 1986)	Read sentences by male and female speakers with various dialects. Transcribed into phonemes and words.

[2] https://huggingface.co/facebook/wav2vec2-large-robust-ft-libri-960h

[3] https://librivox.org

[4] https://huggingface.co/patrickvonplaten/unispeech-large-1500h-cv-timit

Scoring Models

Following transcriptions, the transcribed text was submitted to one of two deep learning transformer-based (Vaswani et al., 2017) language models. Transformer-based language models are probabilistic models that approximate the likelihood of a particular sequence of words appearing in a sentence (Ponte & Croft, 1998; Milokov et al., 2013).

The transformer models are trained on large datasets such as Wikipedia or Books Corpus and are fine-tuned to other classification tasks. The Bidirectional Encoder Representations from Transformers (BERT) model demonstrated how pretraining can be effectively leveraged for a range of downstream tasks (Devlin et al., 2018), including the Generalized Language Understanding Evaluation (GLUE) tasks (Wang et al., 2018). These same models were used in the recent National Assessment of Educational Progress automated scoring challenge for reading and were among the grand prize winners.[5]

Specifically, ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately; Clark et al., 2020) architectures and FNet (Lee-Thorp et al., 2021) were used in Autoscore^{SP}. ELECTRA modifies the BERT architecture to use fewer parameters and uses a different training mechanism to better predict masked words. ELECTRA was also trained on publicly available data (Wikipedia and Books Corpus). FNet replaces the perceptron-based attentional elements of the BERT architecture with Fourier transforms, allowing for faster and less memory-intensive calculations. FNet was trained on the combined set of responses that were transcribed by Unispeech. The two transformer models were then fine-tuned to each item's scores. ELECTRA was used with the output from the wav2vec 2.0 transcriber, and FNet was used with the output of the Unispeech transcriber. Prior to submitting the transcribed responses to the ELECTRA engine, the responses were spell-corrected. The responses submitted to FNet were not spell-corrected because this model was built on transcribed (and un-spell-corrected) responses.

Transformer models offer several key benefits in scoring responses. As noted earlier, such models are trained to understand word usage in context (e.g., prediction of masked words) on very large datasets using millions of parameters. This approach means that transformer models can represent many patterns within and across responses and reflect the variety of language around a particular task (Monarch, 2021). Second, these models use a tokenization scheme that ensures that all words in a response are used—albeit sometimes converted to subwords (i.e., commonly-occurring segments of words) or characters. This strategy ensures that words with unusual spellings are not ignored and are still used to predict scores. Finally, the models are fine-tuned to a scoring task—that is, model parameters are adjusted via backpropagation to maximize scoring accuracy. In essence, fine-tuning adapts the language model to the characteristics of training responses and scores for a given item.

While ELECTRA and FNet were used to model the transcribed test to predict scores, Autoscore^{SP} can be modified to include a range of deep-learning models and any combination of transcribers and scoring models.

Ensembler

Because the models are trained on different data sources and architectures, they should be modeling different aspects of language associated with scores. This allows them to be effectively ensembled, since ensembled models tend to produce higher accuracy than individual models in this scenario (Zhou et al., 2002; Wolpert, 1992). The ensemble model uses a one-versus-rest logistic regression model using the outputs of the individual models (logits) as inputs.

[5] info/results.md at main · NAEP-AS-Challenge/info · GitHub

Condition Codes

Three condition codes were used when modeling and scoring responses (refer to Table 2). The 'No Response' condition code indicates that the transcribers produced blank transcriptions. The 'Low Transcriber Confidence' code indicates that the transcriber had very little confidence in its transcription; such responses can be routed for human review. The 'Not Enough Data' code was used when the transcription produced only one character; such a short response should not be eligible for any score above 1 on the rubric.

Table 2. Condition Codes Assigned by Autoscore^{SP}

Condition Code	Description	Threshold
No Response	Response is blank after either transcription.	No threshold
Low Transcriber Confidence	Wav2vec 2.0 transcriber produced low confidence values on average.	Average confidence value less than .50
Not Enough Data	Wav2vec 2.0 transcribed response is extremely short.	Contains one character

Method

In this section we describe the program and items, data, hand-scoring, and methods used to train and evaluate engine performance.

Program and Items

The English Language Proficiency (ELP) testing program supports educators, states, and members of the public as they implement the ELP Standards and college- and career-readiness standards (CCSSO, 2014). The screener test is used to identify students who qualify for English language development services and is an assessment of a student's language proficiency in the domains of listening, reading, writing, and speaking. The assessment system includes tests on listening, reading, speaking, and writing for students in kindergarten, grade 1, grades 2–3, grades 4–5, grades 6–8, and grades 9–12.

Fifteen speaking items were analyzed in the study. All were single-interaction items with maximum score points ranging from 3 to 5. The minimum score point for all items is 0 (refer to Table 3).

Table 3. Items

Grade Band	Item ID	Task Type	Number of Interactions	Min Score Pts	Max Score Pts
1	694	Picture Description	1	0	3
2–3	1378	Compare Pictures	1	0	3
4–5	2078	Analyze a Visual	1	0	5
4–5	2080	Analyze a Visual	1	0	5
4–5	2108	Compare Pictures	1	0	3
6–8	2662	Analyze a Visual and a Claim	1	0	5
6–8	2664	Analyze a Visual and a Claim	1	0	5
6–8	2694	Language Arts Presentation	1	0	3
6–8	2696	Language Arts Presentation	1	0	3
6–8	2698	Language Arts Presentation	1	0	3
9–12	3344	Analyze a Visual and a Claim	1	0	5
9–12	3346	Analyze a Visual and a Claim	1	0	5
9–12	3400	Language Arts Presentation	1	0	3
9–12	3402	Language Arts Presentation	1	0	3
9–12	3404	Language Arts Presentation	1	0	3

The speaking rubrics varied by task type and are described in Table 4. The rubrics examine the degree to which the test taker uses accurate grammar and word choice and the degree to which the test taker provides a meaningful response to the task demands. The rubrics do not assess verbal fluency and pronunciation. Across all tasks, responses receiving a score of 0 failed to address the communicative demands of the task, did not use English, indicated a refusal to respond, were unintelligible, or were off task/topic. Responses received a condition code if they were blank (coded as A) or had a technological issue impeding the ability to understand the audio (coded as B).

Data

Data were drawn from two U.S. states during the academic years of 2017–2018 and 2018–2019. Data were used in two ways: to obtain transcriptions of student audio and to train and validate the scoring engine. The data were sampled independently for each purpose. For transcriptions, 500 responses were randomly drawn for each item. For engine training and validation, 3,000 responses were drawn for each item. If 3,000 responses were not available, then all available responses were used.

Table 4. ELPA21 Screener Rubric Descriptions

Task Type	Rubric Description
Compare Pictures / Picture Description	Highest scoring responses use appropriate and relevant vocabulary and effective grammatical structures. They describe the main features of the pictures. Lower scoring responses use limited phrases or their grammar/vocabulary use interferes with meaning.
Analyze a Visual (and Variations)	Highest scoring responses use largely accurate grammar and vocabulary/word choice such that any errors do not interfere with the test taker's intended meaning. These responses are well-developed and supported. Lower scoring responses may not address the task or may use grammar or words that interfere with the intended meaning or may use isolated English phrases or repeat the prompt or directions.
Language Arts Presentation	Highest scoring responses use largely accurate grammar and vocabulary/word choice such that any errors do not interfere with the test taker's intended meaning. These responses are well-developed and supported. Lower scoring responses may not address the task or may use grammar or words that interfere with the intended meaning, or use isolated English phrases.

Transcription

Transcriptions were conducted by Measurement Incorporated in fall, 2020. Each student's audio was transcribed by one trained transcriber and the transcription was then reviewed and edited by a second transcriber. This second transcription served as the final transcription. Portions of the audio containing unintelligible words or phrases were identified by tags within the body of the transcription (refer to Table 5). Responses that were entirely blank were flagged as blank.

Table 5. Transcription Tags

Tag	Tag Description
<unintelligible/>	Words or phrases could not be discerned.
<background/>	Words or phrases are from background voices.
<nonenglish/>	Words or phrases are not in English.

Other rules for transcriptions specified including end-of-sentence punctuation but not mid-sentence punctuation and providing correct spelling for words, even with unusual pronunciation (Bawston vs. Boston).

Hand-scoring

Once the random sample of 3,000 test taker records was drawn, the responses and first human score assigned during the original test administration were submitted to Measurement Incorporated for further hand-scoring. These data received a second, independent score, and any response in which the submitted first score and the second score differed was routed for expert adjudication. This hand-scoring was conducted in spring, 2022.

Evaluation Metrics

Evaluation metrics included measures of transcription quality and measures of score quality.

Transcription Metrics

The similarity of the engine transcriptions to the human transcriptions was examined using the Word Error Rate (WER). In the computation, one transcription is chosen as the target (the human transcription), and other is designated as the hypothesis (the engine transcription). The WER calculation uses the Levenshtein distance, which is a method for computing the minimum number of word-level changes (insertions, deletions, substitutions) needed to convert the hypothesis to the target. This value is divided by the number of words in the target text to normalize the metric. In the WER calculations for this study, case is ignored and any punctuation is removed. WER ranges from 0 (the two transcriptions match exactly) and 1 (no engine-transcribed words match the human transcriptions). Note that any word-level differences are counted as differences, including minor misspellings or changes to word endings (e.g., plural versus singular). WER is calculated at the response level and the average is computed across responses.

In addition to the WER for student responses to the 15 test items, we also provide WER for the two transcribers for publicly available datasets. These results illustrate how the transcribers perform on adult speech that is similar to the data on which they were trained.

Scoring Metrics

The engine performance on the validation sample was examined relative to hand-scoring. Quadratic weighted kappa (QWK) and exact agreement statistics were used to examine the level of agreement between the engine scores and the final resolved score. These values were also computed on the two independent human rater scores. The standardized mean differences (SMDs; using the pooled standard deviations) were calculated between the engine scores and the final human scores.

A subset of thresholds from Williamson et al. (2012) was used to highlight any differences. Namely, any item with a final score-Autoscore^{SP} QWK lower than .1 of that calculated on the two humans was noted, as was any standardized mean difference absolute value exceeding .15 (refer to Table 6).

Table 6. Flagging Criteria

Flag	Description	Threshold
QWK	QWK of final score-Autoscore^{SP} minus QWK of two humans	Below -0.10
SMD	Magnitude of the standardized mean difference of the Autoscore^{SP} and final score	Greater than 0.15

Training and Validation Methods

Models were trained for each item. The automated scoring engine was trained on the final resolved score. The 3,000 responses were divided into the training (70%), ensemble (15%), and held-out validation (15%) samples using a stratified random sampling on the final resolved score. The stratification approach ensured that the score distributions across the three samples was similar. The training data were used to train each of the two scoring models. The ensemble data were used to estimate the logistic regression parameters for the ensemble. Finally, the held-out validation data were used to evaluate the engine performance.

Deep learning networks are optimized using methods derived from stochastic gradient descent, which iteratively estimates the gradient

using training batches to find the minimum of a loss function. The number of epochs determines how many times the neural network sees the entire dataset. The learning rate determines the size of the step taken in the optimization method. The batch size determines the number of training samples used to approximate the gradient vector. During training, the number of epochs, the learning rate, and the batch size were examined using an exhaustive grid search with the best performing set of parameters across items chosen for use in final modeling.

Results

This section presents the transcription results and the scoring results, including condition code agreements.

Transcription Results

The transcription results on the publicly available datasets are presented first, followed by the transcription results on the student response data.

Publicly Available Datasets

The average WER for three datasets was computed for Unispeech and for wav2vec 2.0. Recall that wav2vec 2.0 was fine-tuned on the LibriSpeech training data and so is optimized for that dataset. Also, note that Unispeech was pretrained to model phonemes in its latent speech representations explicitly and was fine-tuned on the TIMIT training dataset. As can be observed in Table 7, wav2vec 2 WER values ranged from 7% (LibriSpeech) to 23% (Common Voice), and Unispeech WER values were much higher, ranging from 25% (TIMIT) to 67% (Common Voice). Both transcribers showed the best relative performance on the dataset in which they were fine-tuned. And, both transcribers performed the worst on the Common Voice dataset. Finally, note that these values are higher than those reported in the research for these transcribers; this is because that research conducted post-processing on the transcriptions to enhance the match. We did not do this because we round that the post-processing on the student results tended to result in worse overall performance. Thus, we present the non-post-processed results here.

Table 7. Word Error Rate for Publicly Available Test Datasets

Source	N	Unispeech	Wav2vec 2.0
LibriSpeech	2939	47%	7%
Common Voice	16150	67%	23%
TIMIT	1680	24%	9%

Note. Blank transcriptions removed from sample.

To illustrate WER, Table 8 presents the original human transcription alongside the Unispeech and wav2vec 2.0 transcriptions (and WER values) for a small sample from the Common Voice dataset. We chose to illustrate these data because neither of the transcribers were fine-tuned on these data and the dataset includes speakers with accents. Thus, the transcriptions are likely to be more similar to those in the screener data than the other datasets. The wav2vec 2.0 transcriptions more closely match the human transcriptions but do mis-transcribe some words. The Unispeech transcription appears to have a strong phonetic component, showing greater phonetically-driven mis-spellings.

While the Unispeech clearly does not perform well on the transcriptions as measured by the WER and via visual inspection, we included it in the modeling as a secondary transcription approach because of its focus on phonetic speech and because its associated language model also used phonetic inputs.

Table 8. Selected Example Unispeech and Wav2vec 2.0 Transcriptions from Common Voice

Human Transcription	Unispeech		Wav2vec 2.0	
	Transcribed Text	WER	Transcribed Text	WER
all rings are rngs	e aoll rangs are ronsec	100%	all the rings are rons	50%
decreases in air pressure occurred throughout the bahamas providing strong indications of a disturbance	decreses win arpressure occured thoroughout the bomas providing strong indigations of ar distourbense	71%	decreases in air pressure occurred throughout the bamas providing strong indications of a disturbance	7%
all of the organs are exposed in our exhibition hall	all of the ordgens ar exposed inour exibision hullee	60%	all of the argents are exposed in our exhibition hall	10%
she assisted in the creation of metropol	e eshe asosted inthe krasionof metradol	100%	she assisted in the creation of metra bort	29%
several plaques commemorating the loss of the angel gabriel have been placed near pemaquid	eseveral placks commemerating the los of the angel gabreal havfe ben playstd nea pemoqwid	71%	several plaks commemorating the loss of the angel gabriel have been placed near pemaquid	7%
find me animated movies at amco entertainment tomorrow	he lfind m animated movies at amkho entertainment tomorro	63%	find me animated movies at ampco entertainment to morrow	38%
church and state are separated for a reason	echurch and stat are seperated for a reason	25%	church and state are separated for reason	25%
adriana spoke portuguese fluently	adrana spoe portuges fluentlleeeeeee	100%	adriana spoke portuguese fluently	0%

Response Data

Recall that 500 responses were randomly selected from the response dataset and routed for human transcription. Responses with unintelligible audio, background voices, or non-English had flags inserted to represent the segment of audio with that description. For each item, most responses contained no transcription flags (refer to Table 9). The unintelligible flag was the most common flag, followed closely by the background voices flag. Very few responses had non-English flags. The lower grades (1, 2–3, 4–5) tended to have proportionally more unintelligible flags, and the percentage of background voice flags was consistent across grades and items.

Table 9. Transcription Flags

Grade	Item ID	N	Contains No Flags		Contains Unintelligible Flag		Contains Background Voices Flag		Contains non-English Flag	
			N	%	N	%	N	%	N	%
1	694	477	358	75%	81	17%	38	8%	11	2%
2–3	1378	483	373	77%	76	16%	39	8%	6	1%
4–5	2078	487	437	90%	27	6%	25	5%	3	1%
4–5	2080	488	411	84%	53	11%	33	7%	5	1%
4–5	2108	498	393	79%	78	16%	38	8%	0	0%
6–8	2662	494	443	90%	26	5%	27	5%	1	0%
6–8	2664	490	423	86%	35	7%	37	8%	1	0%
6–8	2694	493	405	82%	50	10%	42	9%	3	1%
6–8	2696	491	441	90%	23	5%	31	6%	1	0%
6–8	2698	490	414	84%	39	8%	45	9%	0	0%
9–12	3344	493	449	91%	18	4%	32	6%	0	0%
9–12	3346	489	414	85%	34	7%	45	9%	1	0%
9–12	3400	492	439	89%	25	5%	31	6%	1	0%
9–12	3402	495	436	88%	21	4%	40	8%	3	1%
9–12	3404	494	417	84%	39	8%	46	9%	2	0%
	Average	**7354**	**6253**	**85%**	**625**	**8%**	**549**	**7%**	**38**	**1%**

Note. Blank responses or responses containing only transcription flags were removed from sample.

The WER error rate was calculated overall, and for responses containing transcription flags and with no transcription flags (refer to Table 10). As with the publicly available datasets, the Unispeech WER was higher than the wav2vec 2.0, with Unispeech having an average value of 64% across items and responses and wav2vec 2.0 having a value of 27% for responses with no transcription flags. The WERs for both transcribers tended to decrease as the grade level increased, suggesting that older student audio is easier for the models to accurately transcribe. This result is not surprising given that the models are trained on adult speech. The responses containing transcription flags show much higher WERs for both transcribers, suggesting that when the humans indicate difficulty transcribing portions of the audio, the transcribers tend to struggle with the audio, as well. Finally, the WER values for all transcriptions were slightly higher (4-5%) than the responses with no transcription flags.

Table 10. Word Error Rate for Responses Containing No or at Least One Human-Assigned Flag in Transcription

Grade Span	Item ID	No Human-Assigned Flags In Transcription			Human-Assigned Flags In Transcription			All Transcriptions		
		N	US	W2V	N	US	W2V	N	US	W2V
1	694	358	71%	40%	119	94%	85%	477	77%	51%
2–3	1378	373	71%	34%	110	87%	83%	483	75%	45%
4–5	2078	437	66%	34%	50	83%	68%	487	68%	37%
4–5	2080	411	68%	31%	77	90%	74%	488	71%	38%
4–5	2108	393	68%	26%	105	81%	47%	498	71%	30%
6–8	2662	443	69%	32%	51	80%	76%	494	70%	37%
6–8	2664	423	64%	27%	67	82%	62%	490	66%	32%
6–8	2694	405	64%	24%	88	86%	59%	493	68%	30%
6–8	2696	441	66%	22%	50	97%	76%	491	69%	27%
6–8	2698	414	63%	22%	76	79%	52%	490	65%	26%
9–12	3344	449	73%	37%	44	98%	84%	493	75%	42%
9–12	3346	414	68%	29%	75	81%	50%	489	70%	32%
9–12	3400	439	49%	14%	53	71%	39%	492	51%	17%
9–12	3402	436	53%	16%	59	84%	59%	495	57%	21%
9–12	3404	417	53%	16%	77	76%	47%	494	56%	20%
	Average	**6253**	**64%**	**27%**	**1101**	**84%**	**64%**	**7354**	**67%**	**32%**

Scoring Results

Next, we present the scoring results on the held-out validation sample, starting with condition code assignments and then score assignments. We also include the performance of the individual engines to illustrate the quality of their scoring and whether ensembling produced any benefit.

Condition Code Assignment

The human raters assigned 195 condition codes in the held-out validation sample across items, and the engine assigned 180 condition codes. As noted in Table 11, 82.6% (n=161) of the human-assigned condition were also assigned an engine condition code. Of the 34 responses that received human-assigned condition codes but were assigned scores by the engine, 26 received scores of 0, 7 received scores of 1, and 1 received a score of 3. Of the 19 responses that received engine-assigned condition codes but were assigned scores by the human raters, 18 received scores of 0 and 1 received a score of 1 by the human raters.

Table 11. Agreement of Engine-Assigned Condition Codes and Human Codes/Scores Across All Items

	Human Scores/Codes				
Engine Scores/Codes	**A**	**B**	**0**	**1 or Higher**	**Total**
No Response	135	0	4	1	140
Low Transcriber Confidence	6	0	4	0	10
Not Enough Data	20	0	10	0	30
0	25	1	NA	NA	26
1 or Higher	6	2	NA	NA	8
Total	**192**	**3**	**18**	**1**	

QWK, Exact Agreements, and SMDs

The QWK, exact agreements, and SMDs are presented in Table 12 for the ensembled model and Table 13 for the individual models. Means, standard deviations, and score point distributions appear in Table 14 for the ensembled model and Table 15 for the individual models.

For all items, the HSAS (final resolved score and Autoscore^{SP} score) QWK was within .1 of the H1H2 (two human rater) QWK values. The HSAS QWK was higher than the H1H2 QWK for 11 of the 15 items; most items with lower QWKs were at the lower grades. The HSAS exact agreement rates were higher than the H1H2 agreement rates for all items but one (grade 1 item). Averaged across items, the HSAS QWK was .84 compared to the .81 for H1H2. The average HSAS exact agreement percentage was 75% versus 68% for H1H2. We do expect the HSAS agreements to be higher than the H1H2 agreements because the engine is both trained on the final resolved score and because the final resolved score is more reliable than any individual human rater score, having gone through a resolution process. Finally, recall that the second human score and resolutions were assigned at a different time period than the first human score was assigned.

Table 13 presents the QWK and exact agreement results for Unispeech-FNet (US) and wav2vec 2.0-ELECTRA (W2V) based models on the held-out validation sample as well as the ensemble agreements from Table 12 (HSAS). The results show that the wav2vec 2.0-based model outperforms the Unispeech-based model, and that the wav2vec 2.0-based model closely mirrored the ensembled model results. These results suggest that the Unispeech-FNet model did not add value in the Ensembling in terms of agreement.

Table 12. Engine-Human QWK and Exact Agreements on the Held-Out Validation Sample

Grade Span	Item ID	N	QWK			Exact Agreement		
			H1H2	HSAS	Diff	H1H2	HSAS	Diff
1	694	391	0.88	0.82	-0.06	81%	74%	-7%
2–3	1378	428	0.87	0.89	0.02	74%	79%	5%
4–5	2078	438	0.90	0.88	-0.02	75%	76%	1%
4–5	2080	438	0.84	0.83	-0.01	61%	68%	7%
4–5	2108	439	0.68	0.78	0.10	67%	79%	11%
6–8	2662	437	0.87	0.92	0.05	62%	75%	14%
6–8	2664	426	0.87	0.93	0.06	57%	74%	17%
6–8	2694	435	0.76	0.80	0.04	66%	73%	7%
6–8	2696	429	0.76	0.75	-0.01	65%	68%	3%
6–8	2698	434	0.83	0.83	0.01	74%	78%	4%
9–12	3344	433	0.90	0.91	0.01	71%	76%	5%
9–12	3346	429	0.88	0.89	0.02	63%	70%	7%
9–12	3400	439	0.74	0.78	0.05	71%	79%	8%
9–12	3402	441	0.69	0.83	0.14	62%	81%	19%
9–12	3404	440	0.73	0.77	0.04	70%	77%	7%
Average			**0.81**	**0.84**	**0.03**	**68%**	**75%**	**7%**

Note. H1H2 = human rater 1 and human rater 2 agreement. HSAS = Final resolved score and Autoscore^{SP} agreement.

Table 13. Ensemble, Unispeech-FNet, and Wav2vec 2.0-ELECTRA and Human QWK and Exact Agreements on the Held-Out Validation Sample

Grade Span	Item ID	N	QWK			Exact Agreement		
			HSAS	US	W2V	HSAS	US	W2V
1	694	391	0.82	0.75	0.85	74%	69%	74%
2–3	1378	428	0.89	0.87	0.91	79%	77%	80%
4–5	2078	438	0.88	0.84	0.90	76%	69%	77%
4–5	2080	438	0.83	0.80	0.84	68%	59%	68%
4–5	2108	439	0.78	0.77	0.75	79%	78%	77%
6–8	2662	437	0.92	0.87	0.91	75%	69%	73%
6–8	2664	426	0.93	0.87	0.92	74%	61%	73%
6–8	2694	435	0.80	0.74	0.81	73%	65%	73%
6–8	2696	429	0.75	0.73	0.74	68%	65%	65%
6–8	2698	434	0.83	0.76	0.82	78%	69%	77%
9–12	3344	433	0.91	0.89	0.91	76%	66%	74%
9–12	3346	429	0.89	0.86	0.89	70%	62%	69%
9–12	3400	439	0.78	0.73	0.78	79%	75%	79%
9–12	3402	441	0.83	0.66	0.82	81%	70%	81%
9–12	3404	440	0.77	0.77	0.76	77%	77%	77%
Average			**0.84**	**0.79**	**0.84**	**75%**	**69%**	**75%**

Note. HSAS = Final resolved score and Autoscore^{SP} agreement. US = Unispeech – FNet and final, resolved score. W2V = wav2vec 2.0 – ELECTRA and final, resolved score.

For all items, the HSAS SMD lies within .15. For 12 of the 15 items, the SMD is positive, indicating that the Autoscore^{SP} engine assigns slightly higher scores than the final, resolved score. This result is likely due to a slight upward bias by the engine, likely due to the distributional characteristics of the final resolved scores. In the table, the H1H2 SMDs are presented, as well, and show substantially more variation than the HSAS scores. This result may be due to differences in rater training or monitoring during the two time periods (H1 during an operational administration; H2 during a follow-up scoring process). Finally, note that the HS and AS standard deviations were similar.

Table 14. Human and Engine Means, Standard Deviations, and Standardized Mean Differences on the Held-Out Validation Sample

Grade Span	Item ID	N	Mean		SD		SMD	
			HS	AS	HS	AS	H1H2	HSAS
1	694	391	2.10	2.15	1.00	0.99	0.00	0.05
2–3	1378	428	1.90	1.97	1.04	1.02	0.00	0.07
4–5	2078	438	2.84	2.91	1.34	1.26	0.03	0.05
4–5	2080	438	2.85	2.78	1.27	1.25	0.17	-0.06
4–5	2108	439	2.46	2.49	0.71	0.72	-0.32	0.03
6–8	2662	437	2.80	2.78	1.45	1.47	0.15	-0.01
6–8	2664	426	2.97	3.00	1.48	1.51	0.14	0.02
6–8	2694	435	2.15	2.24	0.88	0.86	0.04	0.10
6–8	2696	429	2.11	2.09	0.87	0.89	0.17	-0.02
6–8	2698	434	2.24	2.27	0.87	0.85	0.04	0.03
9–12	3344	433	3.60	3.73	1.42	1.44	0.01	0.09
9–12	3346	429	2.97	2.98	1.29	1.30	0.06	0.01
9–12	3400	439	2.53	2.58	0.74	0.79	-0.14	0.07
9–12	3402	441	2.39	2.40	0.77	0.78	-0.07	0.01
9–12	3404	440	2.52	2.61	0.77	0.73	-0.05	0.12

Note. A positive HSAS SMD indicates that the AS mean score is greater than the HS mean score.

Table 15 presents the Unispeech-FNet and wav2vec 2.0-ELECRA based model means, standard deviations, and SMDs on the held-out validation sample. As with the agreement data, both models produce SMD values close to 0, although with some violations of the .15 threshold. The average absolute SMD across items for the ensemble is .05 compared to .09 for the Unispeech-FNet model and .07 for the wav2vec 2.0-ELECTRA model. This result suggests that the ensembling, while not providing improvements in agreements, does provide some stability in producing more similar mean scores compared to the final, resolved score.

Table 15. Human and Unispeech-FNet and Wav2vec 2.0-ELECTRA Means, Standard Deviations, and Standardized Mean Differences on the Held-out Validation Sample

Grade Span	Item ID	N	Mean		SD		SMD	
			US	W2V	US	W2V	US	W2V
1	694	391	2.24	2.10	0.95	1.04	0.14	0.00
2–3	1378	428	1.99	2.00	1.00	1.03	0.09	0.09
4–5	2078	438	2.95	2.86	1.26	1.32	0.09	0.02
4–5	2080	438	2.92	2.79	1.25	1.24	0.06	-0.05
4–5	2108	439	2.52	2.54	0.72	0.70	0.08	0.10
6–8	2662	437	2.86	2.72	1.36	1.47	0.04	-0.05
6–8	2664	426	3.10	3.03	1.41	1.50	0.09	0.04
6–8	2694	435	2.31	2.25	0.91	0.86	0.18	0.12
6–8	2696	429	2.02	2.03	0.85	0.92	-0.10	-0.09
6–8	2698	434	2.32	2.26	0.86	0.86	0.10	0.02
9–12	3344	433	3.73	3.75	1.40	1.47	0.09	0.10
9–12	3346	429	3.08	3.02	1.29	1.30	0.09	0.04
9–12	3400	439	2.60	2.59	0.72	0.73	0.09	0.08
9–12	3402	441	2.41	2.41	0.85	0.76	0.01	0.02
9–12	3404	440	2.63	2.63	0.68	0.68	0.16	0.16

Note. A positive US or W2V SMD indicates that the US or W2V mean score is greater than the HS mean score.

Summary

This report presents the methods and results of CAI's automated scoring engine, Autoscore^{SP}, on 15 English Language Learner (ELL) screener speaking items. The purpose of using automated scoring is to provide scores more quickly so that students can be identified and placed into English language programs sooner and to reduce scoring costs. A random sample of 3,000 responses from two U.S. states from the 2017–2018 and 2018–2019 academic years was drawn, and these responses were hand-scored with any discrepant response receiving an adjudicated read. The performance of the engine was examined on a held-out validation sample (15%). In addition, a separate, smaller sample of responses to each item (n=500) was transcribed by humans to evaluate the engine transcriptions.

Autoscore^{SP} scores student speech by first preprocessing the response audio and then sending the processed response to two transcribers (Unispeech and wav2vec 2.0), the transcribed output to two score prediction models (FNet and ELECTRA), and then to an ensemble that combines with two outputs to predict a final score. The Unispeech word error rate (67%) was higher than the wav2vec 2.0 (32%), averaged across items and for all responses. These values were similar to those observed in the Common Voice dataset. The WERs for both transcribers tended to decrease as grade level increased, suggesting that older student audio is easier for the models to accurately transcribe. This result is not surprising given that the models are trained on adult speech.

Averaged across items, Autoscore^{SP} had a slightly higher QWK (.84) compared to humans (.81). There were no QWK values lower than .10 of human QWK values and no SMD values exceeded .15. In terms of QWK and exact agreement, the Autoscore^{SP} aggregate results were similar to the wav2vec 2.0-ELECRA model but were higher than the Unispeech-FNet model. In terms of SMD, the Autoscore^{SP} results were better than any individual model, showing less variation across items. These results suggest that Autoscore^{SP} was able to mimic hand-scoring. The results also illustrate that the condition codes used by the engine have reasonable accuracy with the human codes (83%), with the majority of scores assigned when condition codes were not matched as 0 scores.

While Autoscore^{SP} performed well relative to the handscorers, the results indicate that the Wav2vec 2.0 transcriber performed much better than the Unispeech transcriber. Additionally, the Wav2vec 2.0/ELECTRA model outperformed the Unispeech/FNet models and matched the ensemble model performance. This result suggest that another transcriber/model combination should be used or that the Word2Vec 2.0/ELECTRA model be used as the sole scorer. We have found that ensembling two models has generally improved performance over individual models, when models themselves perform well and are sufficiently different in feature use and architecture. Thus, next steps will be to replace Unispeech with anther transcriber in the next version of Autoscore^{SP}, as the architecture has been designed to easily accommodate new transcribers and language models and many other transcribers are available for use. Finally, further work needs to be conducted to examine Autoscore^{SP} for any areas of bias for key ELL subgroups (e.g., gender, race/ethnicity, language of origin).

References

Ardila, R., Branson, B., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F., & Weber, G. (2020). Common Voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2vec2.0: *A framework for self-supervised learning of speech representations.* arXiv. https://arxiv.org/abs/2006.11477v2

Boyer, M. (2020). *Evaluating raters and scores in hybrid scoring systems: A proposal for an operational evaluation framework.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Chen, L., Zechner, K., Yoon, S-Y., Evanini, K., Wang, X., ...Gyawali, B. (2018). *Automated scoring of nonnative speech using Speechrater^{SM} v. 5.0 Engine*. ETS Research Report Series, 2018(1), 1–31.

Cheng, J., D'Antilio, Y-Z., Chen, X., & Bernstein, J. (2014). Automated assessment of the speech of young English learners. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, Baltimore, MD.

Cieri, C., Graff, D., Kimball, O., Miller, D., & Walker, K. (2004, 2005). *Fisher English Training Speech Part 1 Transcripts LDC2004T19*. Philadelphia: Linguistic Data Consortium.

Clark, K., Luong, M-T., Le, Q., & Manning, C. (2020). ELECTRA: *Pre-training text encoders as discriminators rather than generators*. arXiv. https://doi.org/10.48550/arXiv.2003.10555

Council of Chief State School Officers (2014). *English Language Proficiency (ELP) standards, with correspondences to K–12 English Language Arts (ELA), mathematics, and science practices, K–12 ELA standards, and 6–12 literacy standards*. Retrieved from CCSSO website: https://ccsso.org/sites/default/files/2017-11/Final%204_30%20ELPA21%20Standards%281%29.pdf

Devlin, J., Chang, M-W., Lee, K., & Toutanova, K. (2018). BERT: *Pre-training of deep bidirectional transformers for language understanding*. arXiv. https://doi.org/10.48550/arXiv.1810.04805

Foltz, P.W., Yan, D., & Rupp, A.A. (2020). The past, present, and future of automated scoring. In D. Yan, A. Rupp, & P.W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 1–9). Boca Raton, FL: CRC Press.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., & Zue, V. (1986). *The DARPA TIMT acoustic-phonetic continuous speech corpus*.

Ghosh, D., Klebanov, B., & Song, Y. (2020). An exploratory study of argumentative writing by young students: A transformer-based approach. In *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications*, Seattle, WA.

Godfrey, J. & Holliman, E. (1993). *Switchboard-1 Release 2 LDC97S62*. Linguistic Data Consortium, Philadelphia, PA.

Graves, A., Fernandez, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labeling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA.

Habermehl, K., Nagarajan, A., & Dooley, S. (2020). A seamless integration of human and automated scoring. In D. Yan, A. Rupp, & P.W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 286–282). Boca Raton, FL: CRC Press.

Hsu, W-N., Sriram, A., Baevski, A., Likhomanenko, T., Xu, Q., Pratap, V., Kahn, J., Lee, A., Collobert, R., Synnaeve, G., & Auli, M. (2021). *Robust wav2vec 2.0: Analyzing domain shift in self-supervised pretraining*. arXiv. https://arxiv.org/abs/2104.01027

Kahn, J., Riviere, M., Zhen, W., Kharitonov, E., Xu, Q., Mazare, P-E., Karadayi, J., Litchinsky, V., Collobert, R., Fuegen, C., Likhomanenko, T., Synnaeve, G., Joulin, A., Mohamed, A., & Dupoux, E. (2019). *Libri-Light: A benchmark for ASR with limited or no supervision*. arXiv: https://arxiv.org/abs/1912.07875

Lee-Thorp, J., Ainslie, J., Eckstein, I., & Ontanon, S. (2021). *FNet: Mixing Tokens with Fourier Transforms*. arXiv. https://arxiv.org/abs/2105.03824

Loukina, A., Zechner, K., Chen, L., & Heilman, M. (2015). Feature selection for automated speech scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, Denver, CO.

Lottridge, S., & Godek, B. (2022, April). *Examining Hybrid Automated and Human Scoring Results in One State Assessment Program*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Matthias, S., & Bhattacharyya, P. (2020). Can neural networks automatically score essay traits? In *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 85–91), Seattle, WA.

Metallinou, A., & Chen, J. (2014). Using deep neural networks to improve proficiency assessment for children English language learners. In *Proceedings from Interspeech*, Singapore.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv. https://doi.org/10.48550/arXiv.1301.3781

Monarch, R. (2021). *Human-in-the-loop Machine Learning: Active learning and the Annotation of Human-Centered AI*. Manning Publications: Shelter Island, NY.

National Center for Education Statistics (NCES). (2022). *Results from the NAEP automated scoring challenge*. Available at info/results.md at main · NAEP-AS-Challenge/info · GitHub

Ormerod, C., Lottridge, S., Harris, A., Patel, M., van Wamelen, P., Kodeswaran, B., Woolf, S., & Young, M. (2022) Automated short answer scoring using an ensemble of neural networks and latent semantic analysis classifiers. *International Journal of Artificial Intelligence in Education*.

Palermo, C. (2022). *Examining Hybrid Automated/Hand-scoring Results in a Multi-State Design*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

Pearson. (2017). AZELLA Arizona English Language Learner Assessment, 2017 Technical Report. Arizona Department of Education.

Peters, M., Neuman, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations*. arXiv. https://arxiv.org/abs/1802.05365

Ponte, J.M., & Croft, W.B. (1998). A language modelling approach to information retrieval. In *Proc. SIGIR*, pp. 275–281. ACM Press.

Riordan, B., Bichler, S., Bradford, A., Chen, J., Wiley, K., Gerard, L., & Linn, M. (2020). An empirical investigation of neural methods for content scoring of science explanations. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, Seattle, WA.

Rodriguez, P., Jafari, A., & Ormerod, C. (2019). *Language models and automated essay scoring*. arXiv. https://doi.org/10.48550/arXiv.1909.09482

Taghipour, K. & Ng, H-T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1882–1891), Austin, TX.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: *A multi-task benchmark and analysis platform for natural language understanding*. arXiv. https://doi.org/10.48550/arXiv.1804.07461

Wang, C., Wu, Y., Wian, Y., Kumatini, K., Liu, S., Wei, F., Zeng, M., & Huang, X. (2021). Unispeech: Unified Speech Representation Learning with Labeled and Unlabled Data. In *Proceedings of the 38th International Conference on Machine Learning*.

Williamson, D., Xi, X., & Breyer, J. (2012). A framework for the evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31(1)*, 2–13.

Wolpert, D. (1992). Stacked generalization. *Neural Networks, 5(2)*, 241–259.

Zechner, K., Evanini, K., Yoon, S-Y., Davis, L., Wang, X., Chen, L., … Leong, C. W. (2014). Automated scoring of speaking items in an assessment for teachers of English as a foreign language. In proceedings of the ninth workshop on innovative use of NLP for building educational applications (pp. 134–142).

Zechner, K., & Loukina, A. (2020). Automated scoring of extended spontaneous speech. In D. Yan, A. Rupp, & P.W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 368–381). Boca Raton, FL: CRC Press.

Zhou, Z.-H., Wu, J., & Tang, W. (2002). Ensembling neural networks: many could be better than all. *Artificial intelligence*, 239–263.