

Developing and Validating an Automatic Scoring Model for Chatbot-Based Conversational Speech

Danwei Cai, Audrey Kittredge, Ben Naismith, Xiangying Jiang, Kevin Yancey

Duolingo, USA

{danwei, audrey, ben.naismith, xiangying, kyancey}@duolingo.com

Abstract

This paper describes the development and validation of an automated scoring model for open-ended chatbot-based conversational speech among English learners in the Duolingo learning app. The model strongly predicted human ratings, produced reliable scores, and showed concurrent validity with Duolingo English Test speaking items, demonstrating stealth proficiency assessment at scale.

1 Introduction

Automated scoring of free-form speech has grown alongside the increasing use of open-ended speaking tasks in language assessment. This scoring has typically been applied to learners' responses to fixed assessment prompts (Evanini and Zechner, 2019; Naismith et al., 2026), with more recent efforts extending to adaptively sequenced conversational formats such as the Interactive Speaking task in the Duolingo English Test (Park et al., 2026). Automatic scoring of open-ended conversational speech elicited outside of assessment contexts, however, remains largely unaddressed (Litman et al., 2018; Ramanarayanan et al., 2019; McKnight et al., 2023). Since the emergence of generative Large Language Models (LLMs), AI chatbots for conversation practice have become widespread (Wang et al., 2025). AI chatbots embedded in commercial language learning apps (e.g., Babbel, Duolingo) elicit learner speech at scale, making it a promising target for automated scoring. Such scoring could provide stealth assessment of learners' proficiency (i.e., assessment unobtrusively embedded within learning activities) without impacting engagement with the app (Shute and Kim, 2014). In this study, we describe the development and validation of an automated scoring model for open-ended chatbot-based conversational speech among English learners in the Duolingo language learning app.

2 Literature Review

Recent years have seen significant growth in the automated scoring of open-ended speaking tasks in language assessment. Research to date has focused primarily on responses to fixed assessment prompts, where two broad paradigms have emerged. Feature-engineered systems extract linguistic features from Automatic Speech Recognition (ASR) transcripts and acoustic signals, then combine them in models trained to predict human ratings (Higgins et al., 2011; Evanini et al., 2013; Evanini and Zechner, 2019; Naismith et al., 2026). Features typically span subconstructs relevant to communicative competence, such as fluency, pronunciation, lexis, content, grammar, and discourse coherence. End-to-end approaches that predict speaking proficiency directly from speech using self-supervised speech foundation models have also been explored (Qian et al., 2019; Bannò and Matassoni, 2023; Phan et al., 2025), and generally outperform feature-engineered systems on benchmark datasets, though at the cost of interpretability (Cai et al., 2025a). More recently, multimodal LLMs have been explored for rubric-based analytic scoring (Ma et al., 2025; Parikh et al., 2025), though questions remain around consistency, fairness, and construct alignment.

These assessment paradigms have been developed primarily for monologic tasks, in which a single speaker responds to a fixed prompt without an interlocutor. Assessment of speech produced in response to adaptively sequenced prompts delivered in a conversational format with an interlocutor presents distinct challenges beyond those of monologic tasks. Some key differences between speech produced in dialogic versus monologic tasks include opening and closing the conversation, managing the conversation topic and turn-taking, and repairing misunderstandings (see Karatay and Xu, 2025 for an overview). All of these activities place

higher demands on listening, understanding, monitoring, and decision-making (Ahmadi and Sadeghi, 2016). Such dialogic assessment tasks have existed in human-scored contexts, such as the IELTS Speaking test, which involves a structured face-to-face interview with a trained examiner (Read, 2022). The DET Interactive Speaking task represents a recent effort to automate conversational assessment: In an avatar-led adaptive conversation with fixed turns, responses are scored in real time to select follow-up questions (Park et al., 2026).

Scoring open-ended conversational speech in chatbot-based language learning contexts presents unique challenges that are not adequately addressed in the literature (Litman et al., 2018; Ramnarayanan et al., 2019; McKnight et al., 2023). Unlike dialogic tasks designed for assessment, such as the IELTS Speaking test or the DET Interactive Speaking task, AI chatbots for conversation practice are designed for learning (Wang et al., 2025). Automated scoring of speech during these learning activities could provide valuable stealth assessment of proficiency to evaluate skill level, without affecting learner engagement (Shute and Kim, 2014). Chatbots embedded in commercial language learning apps (e.g., Babbel, Duolingo) generate learner speech on a large scale, which makes automated scoring especially relevant.

3 The Present Study

The main purpose of this paper is to address the need for research on scoring open-ended conversational speech. This work was conducted in the context of “Video Call”, a voice-based generative AI chatbot embedded in the Duolingo language learning mobile app (Duolingo Team, 2024). Video Call simulates real-time, open-ended conversation: An avatar greets the learner, proposes a conversation topic, and engages in back-and-forth conversational turns with the learner. Authors (under review) found that after being randomly assigned to use the chatbot for four to seven weeks, learners studying English and Spanish on Duolingo significantly improved their speaking skills on a standardized test, compared to learners who did not use the chatbot.

This paper aimed to answer three research questions:

RQ1. Can a scoring model to assess a language learner’s English proficiency be developed that is (a) based on theoretically motivated

speaking subconstructs, and (b) aligns with human proficiency ratings in the context of a low-stakes AI chatbot? Answering these questions is critical for demonstrating that the model is a valid way to measure the intended construct (American Educational Research Association et al., 2014), in this case, speaking proficiency.

RQ2. Are learners’ scores from the model highly correlated with their scores on a speaking test consisting of DET items, when both scores are obtained within a 2-week window? This would provide “concurrent criterion validity,” i.e., concurrent evidence that the scoring model measures the same construct (speaking proficiency) as a criterion variable (American Educational Research Association et al., 2014). In this case, the criterion is a speaking test consisting of items from the DET, an established test with strong evidence of validity and reliability (Naismith et al., 2026).

RQ3. Are consecutive model scores produced by the same learner within a 2-week window highly correlated? Answering this question provides evidence for “test-retest reliability” of the scoring model over time, as independent Video Calls are randomly equivalent and can be seen as alternate forms of the same “test” (American Educational Research Association et al., 2014).

Below, we describe the development of the scoring model to address RQ1. We then describe a test-criterion study conducted with Duolingo app learners to address RQ2 and RQ3.

4 Video Call Scoring Model Development

This section describes the development of an automated scorer for free-form speaking responses produced in Video Call. We describe the scoring pipeline, the linguistic features, and the human rating dataset used to train and evaluate the model.

4.1 Video Call Scoring Model

As described above, Video Call is a conversational speaking feature embedded in the Duolingo language learning app. Technically, the feature is powered by an LLM for response generation, and learner speech is transcribed using the user client’s on-device ASR engine. The automated scorer predicts a holistic proficiency rating from a learner’s free-form speaking response in a Video Call session. For each session, audio and transcripts are extracted across all conversational turns and concatenated into a single response representation, from

which linguistically grounded features are computed along six dimensions described below.

Content features target aspects of task achievement, especially the degree of development or elaboration in the learner’s response. We approximate this using response length (word, sentence, and character counts), which serves as a simple proxy for how fully a response is developed, but does not directly capture topical relevance.

Discourse Coherence features measure the organization and cohesion of the response, operationalized through three complementary sets of measures (McNamara and Graesser, 2011). Lexical cohesion captures word overlap across sentence pairs along two dimensions: the type of overlap (ranging from exact noun overlap to broader stem and content word overlap) and the scope of comparison (ranging from adjacent to non-adjacent sentence pairs), yielding eight features in total. Semantic cohesion is assessed using Latent Semantic Analysis (LSA) (Landauer et al., 2007), computed as the mean and standard deviation of cosine similarities between sentence-pair representations. Referential cohesion is assessed through coreference overlap, measuring the proportion of pronouns in each sentence that refer to entities introduced in the preceding sentence, averaged over consecutive and all sentence pairs.

Lexis features capture the range and accuracy of the learner’s vocabulary use, operationalized through four feature sets. Differential Use (DU) features (Attali, 2011) capture lexicogrammatical patterns distinguishing high- from low-proficiency learners via log odds from n-gram models trained on learner corpora. Lexical diversity is measured using the Measure of Textual Lexical Diversity (MTLD; McCarthy and Jarvis, 2010). Vocabulary sophistication is assessed through the proportion of words at each Common European Framework of Reference (CEFR; Council of Europe, 2020) level (A2–C2) mapped via an internal dictionary (Xia et al., 2016). A vocabulary control measure evaluates contextual appropriateness of word use via the log-ratio of trigram to unigram probabilities from learner reference corpora (Crossley et al., 2012).

Grammar features measure the complexity and accuracy of the learner’s grammatical production, operationalized through four sets of features. Grammatical accuracy is captured through error rates across 52 error types, detected using a grammatical error correction model (Liu et al., 2020) and classified with ERRANT (Bryant et al., 2017). Following the same approach as the lexical DU fea-

tures (Attali, 2011), n-gram models over part-of-speech sequences and dependency-based representations capture shallow and syntactic grammatical patterns distinguishing high- from low-proficiency learners. Grammatical complexity is approximated through dependency tree depth and mean sentence length (Schwarm and Ostendorf, 2005). Mean word length is included as an additional proxy for morphological complexity (Attali and Burstein, 2004).

Fluency features are derived from ASR timestamps and capture how naturally and continuously the learner maintains speech. We operationalized fluency through three sets of features. Speech rate is measured at both the overall and effective (pause excluded) levels in words and syllables per second (Yan et al., 2025). Pause-based features characterize disfluency patterns through the frequency, duration, and variability of inter-phoneme, inter-clause, and within-clause pauses (Yan et al., 2025). Additionally, an end-to-end neural sequence fluency model provides a holistic estimate derived directly from the speech signal, trained on an internal corpus with human fluency ratings and achieving correlations with human ratings comparable to inter-rater agreement levels.

Pronunciation features assess the intelligibility of the learner’s speech. A single comprehensive pronunciation score, produced by a deep neural network (DNN)-based pronunciation scoring model (Liu et al., 2023; Cai et al., 2025b), is used to assess overall phonological control, sound articulation, and prosodic features. The model was trained on an internal speech corpus with human ratings of pronunciation across CEFR-aligned linguistic competence scales, achieving correlations with human ratings comparable to inter-human agreement levels (Cai et al., 2025b).

In total, 117 numerical features are extracted across these six dimensions. Scoring follows a hierarchical structure: For each dimension, a regression model is trained to aggregate its corresponding features into a subconstruct-level score. A final regression model is then applied across the six subconstruct scores to produce a holistic proficiency estimate for the session. All regression models are trained on a human rating dataset described in the following section using 5-fold cross-validation.¹

¹In 5-fold cross-validation, the dataset is split into 5 equal partitions. In each of 5 rounds, 4 partitions are used for training and the remaining 1 is held out for evaluation. Each session is evaluated once, always on a fold it was never trained

This scoring model shares the same subconstructs as the Duo Speaking Scorer used by the DET, but is distinct in its features and training data (Naismith et al., 2026).

4.2 Human Rating Dataset

To train and evaluate the automated scorer, we constructed a human-rated dataset of conversation recordings from Video Call. We sampled 834 sessions from production data, stratified across course sections aligned to CEFR proficiency levels in the Duolingo app, to ensure broad coverage of proficiency from A1 to B2. To mitigate first-language (L1)-specific bias, we used each learner’s user interface (UI) language as a proxy for L1, and sampled as evenly as possible across represented L1 groups.

We developed a task-specific rating rubric grounded in the CEFR framework, covering nine proficiency levels (Pre-A1, A1, A1+, A2, A2+, B1, B1+, B2, B2+, with numerical scores from 1 to 9)². Each level is described along seven dimensions: content, discourse coherence, vocabulary (range and accuracy), grammar (complexity and accuracy), fluency, pronunciation, and interactional competence (ability to sustain and manage the conversational exchange). In this way, the rubric ensured coverage of the overall speaking proficiency construct.

Four internal raters with a linguistics background scored all 834 sessions independently using the rubric. Two raters initially piloted the rubric, resulting in only minor wording changes. All four raters calibrated initially on 20 sessions, and subsequent calibration meetings took place after every 100 sessions. A randomly selected 20% subsample (168 sessions) was rated by at least two raters. Inter-rater reliability was high: exact agreement was .65, adjacent agreement (within one rating point) was .95, Quadratic Weighted Kappa (QWK) was .95, and Pearson r was .95. These values indicate strong consistency among raters across the nine-point scale. Of the 834 rated sessions, 116 were excluded due to system errors, poor audio quality, insufficient response sample, or non-target-language speech, yielding a final dataset of 718 usable sessions.

on. The reported performance is the average across those 5 held-out folds.

²Rating rubric is available upon request.

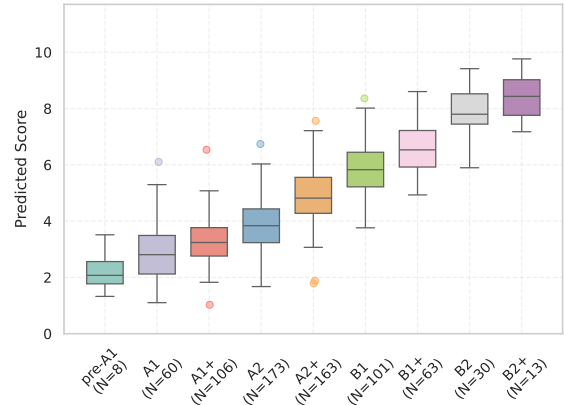


Figure 1: Distribution of automated scorer predictions by human rating category. Each box represents the interquartile range (IQR) of predicted scores for sessions assigned a given CEFR label by human raters; whiskers extend to $1.5 \times \text{IQR}$, and outliers are shown as individual points. Sample sizes per category are shown on the x-axis.

4.3 Model Performance

To evaluate the automated scorer’s ability to predict human judgments of speaking proficiency, we examined the correlation between model predictions and human ratings using 5-fold cross-validation. Model predictions were strongly correlated with human ratings (Pearson $r = .84 \pm .02$; Spearman $\rho = .83 \pm .03$), approaching the inter-rater reliability ceiling of $r = .95$.

Figure 1 shows the distribution of model-predicted scores grouped by human rating category. Data points are drawn from the full dataset, aggregated across all five evaluation folds of the cross-validation, such that each session is represented exactly once using its held-out fold prediction. Predicted scores increase monotonically across CEFR levels from Pre-A1 through B2+, with median predicted scores rising consistently across levels.

5 Test Criterion Study

While the cross-validation results confirm that the model captures the construct as operationalized by human raters, they do not establish whether scores carry meaning beyond the training context. To address this question, this section describes a study conducted to evaluate the concurrent criterion validity and test-retest reliability of proficiency scores produced by the model. We investigated correlations between the model’s scores and scores on speaking items from the DET in the same group of Duolingo app learners, as well as correlations



Figure 2: Speaking test scores plotted against Video Call scores with fitted regression line and 95% confidence interval ($N = 310, \rho = .81$).

between the model’s scores from two consecutive Video Calls made by a subset of these learners.

5.1 Methods

5.1.1 Participants and Procedure

We invited 6251 learners studying at A1 to B2 CEFR proficiency levels in Duolingo English courses for German, Japanese, and Russian speakers to check their eligibility for participation in the study. These learners had access to Video Call through a paid subscription and had used it in the last 24 hours, indicating that they were active users of the feature. They had never taken the DET or the DET practice test to avoid potential prior exposure to items on the speaking test in this study.

Of these learners, 473 confirmed their age as 18 years or older, and were invited to take the speaking test and complete a Video Call in the Duolingo app within one week.

5.1.2 Speaking test

This 15-minute test consisted of four speaking items drawn from the DET practice test, which had previously appeared on the DET. They included speaking responses to written prompts (“Read, then Speak” and “Speaking Sample” items), describing an image prompt (“Speak about the Photo” item), and speaking in response to an auditorily presented prompt (“Listen, then Speak” item) (Naismith et al., 2026). Items appeared in the learner’s UI language (German, Japanese, or Russian). Internal consistency reliability of the test was high (Cronbach’s $\alpha = .96$, Confidence Interval $[.95, .96]$).

Table 1: Correlations of speaking test scores with Video Call scores across course sections ($N = 310$).

Course Section	CEFR	N	Spearman ρ	p
0, 1, 2	A1	31	.77	< .001
3	A2	77	.82	< .001
4	B1.1	36	.83	< .001
5	B1.2	81	.78	< .001
6	B2.1	59	.73	< .001
7	B2.2	26	.69	< .001

Table 2: Correlations of speaking test scores with Video Call scores across UI language groups.

UI language	N	Spearman ρ	p
German	100	.75	< .001
Japanese	161	.71	< .001

5.1.3 Data Analysis

Participants’ responses to individual items on the speaking test were scored using the Duo speaking grader (Naismith et al., 2026), which outputs normalized scores on a standard normal scale, with most values falling in the range of -4 to 4 . Participants’ full speaking test scores were the average of the four speaking test item scores.

Participants’ Video Calls were scored using the model described above, which also outputs normalized scores with most values falling in the range of -4 to 4 . For the purposes of establishing concurrent criterion validity, we only analyzed scores of Video Calls that participants completed within 14 days of taking the speaking test.

We calculated Spearman’s rho correlations between participants’ speaking test scores and Video Call scores, as well as between consecutive Video Call scores of participants who made at least two calls within 14 days of taking the speaking test.

5.2 Results

5.2.1 Concurrent Criterion Validity

There were 310 learners who completed at least one scored Video Call within 14 days of taking the speaking test. The correlation between Video Call scores and speaking test scores was strong, positive, and significant, both overall (Spearman $\rho = .81, p < .001$; see Figure 2) and within course sections (see Table 1), as well as within UI language groups (see Table 2). We did not calculate the correlation for the Russian UI language group, which had fewer than 100 learners, as correlation

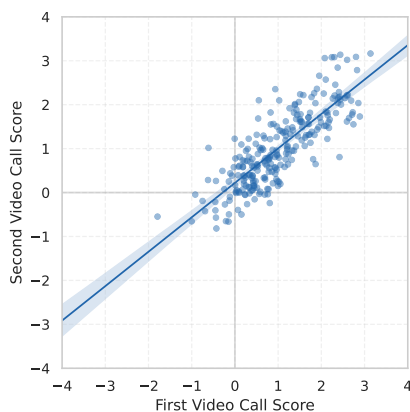


Figure 3: Learners’ first and second Video Call scores with fitted regression line and 95% confidence interval ($N = 210, \rho = .82$).

coefficients can be unstable for such small sample sizes (Schönbrodt and Perugini, 2013). Collectively, these results provide evidence of concurrent criterion validity, showing that the scoring model accurately estimated participants’ speaking proficiency.

5.2.2 Test-Retest Reliability

There were 210 learners who completed at least two scored Video Calls within 14 days of taking the speaking test. The correlation between the two Video Call scores was strong, positive, and significant, both overall (Spearman $\rho = .82, p < .001$; see Figure 3) and for the Japanese UI language group, which had more than 100 learners (Spearman $\rho = .71, p < .001$), suggesting that the model produced reliable scores.

6 General Discussion

The primary purpose of this paper was to describe the development of a model for scoring open-ended conversational speech in English. As with any assessment of speech, whether by humans or an automated scorer, it is also necessary to ensure the reliability, validity, and fairness of the models’ scores (Zechner and Evanini, 2020; Williamson et al., 2012). Specifically, this paper investigated three research questions described in Section 3.

Addressing RQ1, we found that the automated scorer achieved strong agreement with human raters (Pearson $r = .84$, Spearman $\rho = .83$), approaching the inter-rater reliability ceiling of $r = .95$. The human-machine agreement achieved by the scorer is broadly comparable to that reported for feature-engineered systems applied to

prompted speaking tasks in high-stakes tests, which typically falls in the range of .70 – .85 (Educational Testing Service, 2025; Naismith et al., 2026). Even though human scores and machine scores are naturally derived through different means, here the machine scores still reflect a comprehensive model of the speaking construct and relevant sub-constructs. This outcome was achieved because of the model’s research-backed features, motivated by a pre-defined construct, and its training on expert human ratings, which were based on a carefully piloted CEFR-based rubric.

We also found that model scores based on Duolingo app learners’ Video Calls were significantly correlated with the same learners’ scores on an independent speaking test with items from the DET (Naismith et al., 2026), providing initial evidence for the model’s concurrent criterion validity to address RQ2. Furthermore, model scores from consecutive Video Calls made by a subset of these learners were also significantly correlated, providing initial evidence for the model’s test-retest reliability and addressing RQ3.

Collectively, these results provide a proof of concept for scoring open-ended conversational speech from a generative AI chatbot in a mobile language learning app, and anchor the scoring model in rigorous measurement science (American Educational Research Association et al., 2014). Critically, the Duolingo app and others like it are low-stakes in nature. Much of the research on assessing speech is in high-stakes contexts where learners are predominantly motivated to give their maximum effort, where audio quality is often checked and consistent, and where responses are not tied to specific curricular content. These results signal the value of using automated scoring for assessment in a low-stakes language learning app. Moreover, given the wide prevalence of generative AI chatbots in language learning (Wang et al., 2025), these results have significant practical implications. This type of model could support stealth assessment without impacting learners’ engagement (Shute and Kim, 2014), and scores from the model could also be used as formative assessment to enhance personalization of instruction. These benefits may be especially important in the context of large-scale language learning apps, where learner engagement is key to retention and learning.

6.1 Limitations

The scoring model has several limitations. First, content features rely on response length as a proxy for elaboration, and do not capture topical relevance to the conversational exchange. Second, the model does not assess interactional competence. By concatenating all learner turns into a single response representation, it discards the turn-by-turn structure of the conversation, despite interactional competence being a component of the human rating rubric. Third, text-based features, which underlie four of the six subconstructs, are derived from on-device ASR transcripts of variable accuracy. ASR errors propagate into grammar, lexis, and coherence features and may vary systematically across proficiency levels and L1 groups, introducing construct-irrelevant variance. Fourth, the training dataset of 718 sessions is relatively small and skewed toward intermediate proficiency levels, which may limit score precision at the extremes of the scale. Finally, fairness was not formally evaluated: The extent to which scores may vary systematically across L1 groups or demographic subpopulations beyond what is attributable to proficiency differences remains unknown.

The test criterion study also has limitations. Although the speaking test consisted of items drawn from the DET, we did not conduct a detailed psychometric analysis of the 4-item test. Future research should validate the model's scores with more comprehensive tests.

7 Conclusion

This study presents a proof of concept for automated scoring of open-ended conversational speech produced in a generative AI chatbot embedded in the Duolingo language learning app. A feature-engineered scorer extracting 117 linguistically grounded features across six subconstructs was developed. The scorer achieved strong agreement with human raters, approaching the inter-rater reliability ceiling. Evidence of concurrent criterion validity with speaking items from an independent validated test (the DET) demonstrated that the scorer estimates learner speaking proficiency accurately from naturalistic chatbot interactions without disrupting the learning experience. Highly correlated scores from consecutive, independent chatbot interactions also provided initial evidence of the scorer's test-retest reliability. As generative AI chatbots become increasingly prevalent in language learning

contexts, automated scoring of this kind offers a practical pathway to formative stealth proficiency assessment at scale. Future work should investigate how this scoring model can be used to improve app learners' placement, progress monitoring, and personalization of learning experiences.

Acknowledgements

We would like to thank Steven Schwering, Caleb Reed, and Evan Ziebart for help with data collection and analysis, and Ping-Lin Chuang, Hana Ozaki, and Anthony Verardi for help with developing the rating rubric and conducting the human rating study. We would also like to thank Geoff LaFlair, Alina von Davier, Cindy Berger, and Bozena Pajak for comments on a draft version of this paper.

References

- Alireza Ahmadi and Elham Sadeghi. 2016. [Assessing English language learners' oral performance: A comparison of monologue, interview, and group oral test](#). *Language Assessment Quarterly*, 13(4):341–358.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC.
- Yigal Attali. 2011. [A differential word use measure for content analysis in automated essay scoring](#). *ETS Research Report Series*, 2011(2):i–19.
- Yigal Attali and Jill Burstein. 2004. [Automated essay scoring with e-rater® v. 2.0](#). *ETS Research Report Series*, 2004(2):i–21.
- Stefano Bannò and Marco Matassoni. 2023. [Proficiency assessment of L2 spoken English using Wav2Vec 2.0](#). In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1088–1095.
- Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. [Automatic annotation and evaluation of error types for grammatical error correction](#). In *Proceedings of the 55th annual meeting of the association for computational linguistics (Volume 1: Long Papers)*, pages 793–805.
- Danwei Cai, Nitin Madnani, and Kevin Yancey. 2025a. [Team Perezoso's ASR and SLA system for Speak & Improve Challenge 2025](#). In *10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 51–55.
- Danwei Cai, Ben Naismith, Maria Kostromitina, Zhongwei Teng, Kevin P Yancey, and Geoffrey T LaFlair. 2025b. [Developing an automatic pronunciation](#)

- scorer: Aligning speech evaluation models and applied linguistics constructs. *Language Learning*, 75(S1):170–203.
- Council of Europe. 2020. *Common European framework of reference for languages: Learning, teaching, assessment – Companion Volume*. Council of Europe Publishing.
- Scott A Crossley, Zhiqiang Cai, and Danielle S McNamara. 2012. Syntagmatic, paradigmatic, and automatic N-gram. In *Proceedings of the 25th international Florida artificial intelligence research society (FLAIRS) conference*, pages 214–19.
- Duolingo Team. 2024. [How to Video Call on Duolingo](#). Accessed: 2025-10-13.
- Educational Testing Service. 2025. [TOEFL research insight series, volume 3: Reliability and comparability of TOEFL iBT scores](#). Accessed: 2026-04-23.
- Keelan Evanini, Shasha Xie, and Klaus Zechner. 2013. Prompt-based content scoring for automated spoken language assessment. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 157–162.
- Keelan Evanini and Klaus Zechner. 2019. [Overview of automated speech scoring](#). In *Automated Speaking Assessment*, pages 3–20. Routledge.
- Derrick Higgins, Xiaoming Xi, Klaus Zechner, and David Williamson. 2011. [A three-stage approach to the automated scoring of spontaneous spoken responses](#). *Computer Speech & Language*, 25(2):282–306.
- Yasin Karatay and Jing Xu. 2025. [Exploring the potential of conversational AI for assessing second language oral proficiency](#). *TESOL Quarterly*, 59:S220–S250.
- Thomas K Landauer, Danielle S McNamara, Simon Dennis, and Walter Kintsch. 2007. *Handbook of latent semantic analysis*. Psychology Press.
- Diane Litman, Helmer Strik, and Gad S Lim. 2018. [Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities](#). *Language Assessment Quarterly*, 15(3):294–309.
- Wei Liu, Kaiqi Fu, Xiaohai Tian, Shuju Shi, Wei Li, Zejun Ma, and Tan Lee. 2023. [Leveraging phone-level linguistic-acoustic similarity for utterance-level pronunciation scoring](#). In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Rao Ma, Mengjie Qian, Siyuan Tang, Stefano Bannò, Kate M. Knill, and Mark J.F. Gales. 2025. [Assessment of L2 oral proficiency using speech large language models](#). In *Interspeech 2025*, pages 5078–5082.
- Philip M McCarthy and Scott Jarvis. 2010. [MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment](#). *Behavior research methods*, 42(2):381–392.
- Simon W McKnight, Arda Civelekoglu, Mark Gales, Stefano Bannò, Adian Liusie, and Katherine M Knill. 2023. [Automatic assessment of conversational speaking tests](#). In *9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 99–103.
- Danielle McNamara and Arthur C. Graesser. 2011. [Coh-Metrix: An automated tool for theoretical and applied natural language processing](#). In *Applied Natural Language Processing*, pages 188–205. IGI Global.
- Ben Naismith, Ramsey Cardwell, Geoffrey T. LaFlair, Steven Nydick, and Masha Kostromitina. 2026. [Duolingo English Test: Technical manual](#). *Duolingo Research Report*.
- Aditya Kamlesh Parikh, Cristian Tejedor Garcia, Catia Cucchiari, and Helmer Strik. 2025. [Zero-shot speech LLMs for multi-aspect evaluation of L2 speech: Challenges and opportunities](#). In *10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 11–15.
- Yena Park, Yigal Attali, Xiaowan Zhang, Jacqueline Church, Kai-Ling Lo, Andrew Runge, Sarah Goodwin, and Geoffrey T. LaFlair. 2026. [Toward Standardized Interactivity: An AI-Enabled Adaptive Speaking Task for Computer-Delivered Large-Scale Assessment](#). *Language Testing*, 0(0).
- Nhan Phan, Anusha Porwal, Yaroslav Getman, Ekaterina Voskoboinik, Tamás Grósz, and Mikko Kurimo. 2025. [One Whisper to grade them all](#). In *10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 56–60.
- Yao Qian, Patrick Lange, Keelan Evanini, Robert Pugh, Rutuja Ubale, Matthew Mulholland, and Xinhao Wang. 2019. [Neural approaches to automated speech scoring of monologue and dialogue responses](#). In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8112–8116.
- Vikram Ramanarayanan, Keelan Evanini, and Eugene Tsuprun. 2019. [Beyond monologues: Automated processing of conversational speech](#). In *Automated speaking assessment*, pages 176–191. Routledge.
- John Read. 2022. [Test review: The International English Language Testing System \(IELTS\)](#). *Language Testing*, 39(2):297–304.

- Sarah E Schwarm and Mari Ostendorf. 2005. [Reading level assessment using support vector machines and statistical language models](#). In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530.
- Felix D. Schönbrodt and Marco Perugini. 2013. [At what sample size do correlations stabilize?](#) *Journal of Research in Personality*, 47(5):609–612.
- Valerie J. Shute and Yoon Jeon Kim. 2014. [Formative and stealth assessment](#). In *Handbook of Research on Educational Communications and Technology*, pages 311–321. Springer New York.
- Feifei Wang, Alan CK Cheung, Amanda J Neitzel, and Ching Sing Chai. 2025. [Does chatting with chatbots improve language learning performance? A meta-analysis of chatbot-assisted language learning](#). *Review of Educational Research*, 95(4):623–660.
- David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. [A framework for evaluation and use of automated scoring](#). *Educational Measurement: Issues and Practice*, 31(1):2–13.
- Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. [Text readability assessment for second language learners](#). In *Proceedings of the 11th workshop on innovative use of NLP for building educational applications*, pages 12–22.
- Xun Yan, Yuyun Lei, and Yulin Pan. 2025. [Diving deep into the relationship between speech fluency and second language proficiency: A meta-analysis](#). *Language Learning*, 75(4):1051–1090.
- Klaus Zechner and Keelan Evanini, editors. 2020. [Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech](#). Routledge.