



Digital Applied Linguistics, 4, 104860 (2026)
<https://doi.org/10.29140/dal.2026.104860>



Exploring the cross-task transferability of an automated scoring model for discourse completion tasks in relation to learner perceptions

YUKO HAYASHI^a 

YUSUKE KONDO^b 

YUTAKA ISHII^c 

^aFaculty of Education, Saga University, Japan

^bGlobal Education Center, Waseda University, Japan

^cGraduate School of Education, Chiba University, Japan

Abstract

As automated scoring systems are increasingly integrated into technology-mediated speaking practice and assessment, ensuring generalisability and construct alignment across tasks remains a critical challenge, particularly in discourse-based contexts. This study examined cross-task transferability by applying an automated scoring model originally developed for multi-turn discourse completion tasks (DCTs). While prior work demonstrated moderate agreement between human and machine scores within a single task, it remains unclear whether such models maintain reliability when applied to discourse tasks that elicit different linguistic and semantic demands. Using responses from 116 undergraduate EFL learners, the existing random-forest-based model was applied to two oral DCTs without task-specific retraining. Human–machine agreement was examined using multiple agreement indices and confusion matrices, followed by lexical, keyword-based, and regression analyses of learner responses, along with a descriptive analysis of learner perceptions for system evaluation. Results showed high human–human agreement across both tasks, whereas human–machine agreement was considerably stronger for DCT1 than for DCT2. Although task-relevant semantic features were associated with both human and automated scores, particularly for academic-related content in DCT2, confusion matrices and transcript inspection indicated that the automated model did not adequately capture discourse-level relationships reflected in holistic human scoring. These findings suggest that limited cross-task transferability

cannot be attributed simply to insensitivity to task-relevant semantic content but may instead reflect broader challenges in representing discourse contingency within multi-turn interactions. At the same time, learners generally evaluated the automated feedback positively, although positive perceptions should not be taken as evidence of scoring validity. Taken together, the findings highlight the need for greater construct alignment in cross-task automated scoring and point to the potential role of such systems as learning-support tools, rather than fully reliable assessment instruments, in technology-enhanced language learning environments.

Keywords: Automated scoring; discourse completion task; automated feedback, learner perception

Introduction

Owing to rapidly evolving technological innovations in automation and artificial intelligence (AI), automated scoring has become a well-established application of machine learning and is widely used in both high- and low-stakes assessment contexts, including foreign language (FL) and second language (L2) speaking. Most automated speech scoring systems comprise three main components—automatic speech recognition (ASR), feature extraction, and a scoring model (Chen et al., 2018)—and have been applied to tasks ranging from highly constrained to free speech (Zechner & Loukina, 2020).

Highly constrained tasks, such as reading aloud and sentence repetition, provide the advantage of applying consistent criteria and facilitating targeted phonological or phonemic assessment, due to reduced variability in pronunciation and structure (Knoch et al., 2021). However, they do not always reflect a speaker's command of the language in real-life communicative situations, which require more creative and spontaneous responses and often involve communicative strategies such as meaning negotiation. In this regard, free speech may be viewed as a more direct measure of communicative ability. Nevertheless, it presents significant challenges for ensuring rating quality (i.e. reliability and validity) due to the greater unpredictability in responses and variability in content and linguistic features (Norris & Davis, 2024).

This study focused on an oral DCT to address some of these challenges. This task allows the collection of learner responses in a communicative format that involves both comprehension and the production of appropriate and timely responses within a single- or multi-turn context, thereby resembling naturally occurring discourse (Hayashi et al., 2024; Ogiermann, 2018).

Literature Review

Oral Discourse Completion Tasks

Since the 1980s, DCTs have been widely used as data collection methods in various research fields, including cross-cultural pragmatics (Blum-Kulka, 1982; Blum-Kulka & Olshtain, 1984), intercultural communication strategies (Sweeney & Hua, 2010), and L2/FL speaking assessments (Hayashi et al., 2024; Kondo & Ishii, 2017). A DCT, which can be administered in written or oral form, elicits discourse data by requiring test-takers to respond to written or pre-recorded descriptions of specific social situations (Sweeney & Hua, 2016). Thus, DCTs provide insights into test takers' language use in discourse and indicate their communicative language ability in the target language.

Critics, however, highlight the lack of spontaneity and authenticity in DCTs, arguing that they may not accurately reflect naturally occurring discourse because responses are elicited in experimentally controlled, often single-turn, hypothetical situations. This contrasts with real-life interactions, in which speech acts (e.g., offers, refusals and hesitations) are realised through rapid, tightly coordinated turn-taking (Grainger et al., 2015; Houck & Gass, 1996; Kasper, 2000). However, collecting sufficient amounts of naturally occurring spoken data is not always feasible because of practical constraints (e.g., time and personnel) and a range of social and situational factors (e.g., age, proficiency level, power, and social distance) (Eslami & Mirzaei, 2014; Golato, 2003).

Oral DCTs help address these challenges by allowing the observation of longer, more elaborate responses that contain linguistic forms representative of spoken language, especially compared with written DCTs (Yuan, 2001). Oral DCTs administered in a multi-turn context, in particular, may serve as a more direct measure of authentic language use (Hayashi et al., 2024).

Automating Discourse Completion Tasks

A wide range of linguistic tasks with varying degrees of open-endedness have been automated and operationalised in both low-stakes (e.g., practice tests, out-of-class learning) and high-stakes assessment contexts (e.g., TOEFL iBT, Versant) using Automated Speech Recognition (ASR) systems. For instance, IBM Watson Speech-to-Text is a widely used ASR system integrated into the Watson Cloud Services platform. It has demonstrated reliable performance in recognising speech from both native and non-native English speakers (Czyż et al., 2019; Hayashi et al., 2024). Whisper, developed by OpenAI, is another notable ASR model that has gained attention for its multilingual capabilities and robustness to diverse accents and background noise (McGuire & Larson-Hall, 2025; Radford et al., 2023), though its high computational demands can limit its usability in real-time (online) transcription (Macháček et al., 2023).

Hayashi et al. (2024) have demonstrated the feasibility of using IBM Watson as an ASR system to transcribe learner responses to oral DCTs, thereby enabling automated scoring and feedback provision based on response intelligibility and appropriateness. Building on Kondo & Ishii (2017), who conducted automated scoring of learner responses produced through a single-turn oral DCT, Hayashi et al. (2024) developed a new system that employs a multi-turn oral DCT to collect spoken data that more closely resembles naturally occurring interactions. The results showed that the exact agreement between human and machine scores was moderately high at 72%, comparable to other established systems (e.g., SpeechRater; Zechner & Loukina, 2020). Moreover, the system has the advantage of generating immediate feedback in the form of an automated score accompanied by Can-Do descriptors (i.e., rubrics) within regular classroom environments using readily available devices such as tablets. This design feature makes the system easily implementable in local educational contexts without requiring advanced computational resources or external cloud services. For example, recent experimental studies have shown that ASR-based tools can improve speech features such as lexical diversity, fluency, and pronunciation accuracy in both flipped and in-class contexts (Jiang et al., 2022; Liu et al., 2025), yielding pedagogical benefits such as enhanced learner progress and greater efficiency of formative feedback (Ngadiso et al., 2023). Although the predictive capability of the DCT-based system was supported by Hayashi et al. (2024), the efficacy of feedback from the learner's perspective remains unclear, leaving its usability as a learning aid largely unexplored.

Offering Feedback and its Efficacy in Learning

Learners' perceptions of feedback on their performance in language tests, whether paper-based or online, have been increasingly researched to date within FL/L2 assessment (e.g. Jang, 2009;

Sawaki & Koizumi, 2017; Zhai & Ma, 2022). Learners tend to positively perceive feedback that includes specific types of information, such as (a) sets of Can-Do descriptors corresponding to the attained score (current level) alongside those for the next level, thereby highlighting the gap to be filled (Sawaki & Koizumi, 2017), and (b) a holistic score accompanied by criterion-based descriptors (e.g. vocabulary, organisation), revision suggestions, and learning recommendations (Zhai & Ma, 2022). Conversely, learners may report frustration over generic, indirect feedback that lacked actionable guidance and placed a greater cognitive burden on them (e.g., “Consider revising this sentence” as opposed to “Consider replacing this adjective with *significant* or *vital* to maintain an academic tone”) (Ranalli, 2021; Zhang, 2020). In addition, learners’ engagement with feedback could decrease when there is a time lag between test administration and the receipt of the score and feedback; however, detailed feedback may mitigate this effect (Sawaki & Koizumi, 2017).

Automated feedback may enhance learner perceptions and engagement with feedback. A key advantage of automated feedback lies in its immediacy, meaning learners do not need to wait for score reports to be returned, during which they may lose motivation to use feedback for further study (Sawaki & Koizumi, 2017). Spoken language technology allows learners to practise speaking and receive instant feedback anytime and anywhere, supporting language learning. This technology can enhance learners’ motivation and autonomy by providing tailored and diagnostic feedback (Zechner & Hsieh, 2024), leading to improvements in specific learning outcomes, such as pronunciation (e.g., Cucchiari et al., 2009). Such benefits have also been reported in studies of spontaneous speech tasks. For instance, Gu et al. (2021) found that the majority of learners perceived three types of machine-generated feedback included in the speaking section of the TPO test, a low-stakes practice test for the TOEFL iBT, as helpful: (a) two subscores (Delivery and Language Use), (b) a holistic score for each task, and (c) feedback on specific linguistic features such as fluency, vocabulary, and pronunciation. However, learners also noted a few limitations, including a lack of content-related feedback and difficulty interpreting graphical feedback. Beyond spoken language, insights from Automated Writing Evaluation (AWE) research further suggest that facilitating conditions (i.e. the administrative or technical infrastructure that supports the use of a system), AWE self-efficacy (i.e. learners’ judgments of their capacity to use automated feedback to accomplish a task), and system characteristics are significant determinants of college students’ intentions to utilise automated feedback (Zhai & Ma, 2022).

Although the present study focused on rule-based automated feedback derived from human-designed scoring criteria, it is important to acknowledge the growing body of work on AI-based feedback. Recent studies have shown that AI-driven applications can improve learners’ pronunciation and fluency (Mingyan et al., 2025) and reduce speaking anxiety (Ding & Yusof, 2025). These findings highlight the potential of AI-powered approaches. Nevertheless, they complement rather than replace rule-based systems grounded in rubrics and human scoring standards, which continue to provide interpretable, criterion-referenced feedback that is pedagogically transparent and readily applicable in classroom contexts (Rodrigues et al., 2025; Vieri & Petrea, 2025).

Summary

Recent technological advancements have led to a surge in learning and assessment tools equipped with automated scoring and feedback for learner speech, alongside the automation of a wide range of speaking tasks, from highly constrained formats to free speech. However, both task types present limitations in capturing communicative ability. Highly constrained tasks (e.g., sentence repetition) provide limited evidence of learners’ ability to produce creative,

multi-turn discourse, whereas free speech tasks, whether in low- or high-stakes contexts (e.g., TOEFL iBT), pose substantial implementation challenges due to personnel, technical, and time constraints.

Despite growing interest in automated scoring of spoken language, much remains to be learnt about model performance beyond the task context used for development, particularly in discourse-based assessments where linguistic and semantic demands vary across prompts. Models based on surface linguistic features may perform adequately within a single task but fail to capture task-specific semantic variation across contexts. The present study addresses this gap by examining whether an automated scoring model developed for a multi-turn DCT (Hayashi et al., 2024) maintains reliability across additional DCTs eliciting different communicative content. Rather than proposing a new algorithm, this study evaluates the model's stability, construct alignment, and learner-perceived utility across tasks. By integrating agreement analysis, content profiling, and learner perception data, it contributes to ongoing discussions on model transferability, construct representation, and the implications of task-agnostic scoring approaches.

Method

While previous research demonstrated the technical feasibility of multi-turn DCT-based automated scoring, questions about the reliability and learners' perceptions of the feedback remain unresolved. Therefore, the present study examined (a) the across-task transferability of the automated scoring system developed by Hayashi et al. (2024) and (b) how learners perceive the feedback generated by this system. The specific research questions were as follows:

1. Does the automated scoring system demonstrate transferability across different discourse completion tasks?
2. How do learners perceive the automated feedback provided by the system?

Participants

A total of 118 first-year undergraduate students aged 18–20 from the same university participated in this study. The participants constituted a fresh sample independent of that used in Hayashi et al. (2024). All participants were considered intermediate-level EFL learners, corresponding approximately to B1 on the Common European Framework of Reference for Languages: Learning, teaching, assessment (Council of Europe, 2020), with TOEIC scores ranging from 400 to 550 (ETS, 2025). Due to technical glitches, two participants completed DCT1 but not DCT2 and were therefore excluded from the DCT-response analyses, resulting in a final analytical sample of 116.

Discourse Completion Tasks

Two multi-turn oral DCTs (DCT1 and DCT2) were used. Both were modelled on the single-turn DCTs developed by Kondo & Ishii (2017), with DCT2 adapted from DCT1 used by Hayashi et al. (2024). Multi-turn formats were adopted to enhance authenticity by approximating naturally occurring discourse (see Appendix A for full task details).

Following Hayashi et al. (2024), the DCTs, administered via tablets with headsets, were presented consecutively as a single speaking session, with each turn displayed on a separate screen and completed within a fixed response window (10–20 seconds for stimulus presentation, followed by a 5-second countdown and 30 seconds for responding). Instructions and practice tasks were provided prior to task completion. All stimuli were presented both aurally and in

writing to reduce cognitive load and support speech production (Hayase et al., 2018; Hayashi, 2019), thereby helping to maintain the construct validity of the DCT (Chapelle, 2012).

Scoring Rubrics

Drawing on Hayashi et al. (2024), and consistent with the objective of using DCTs as an index of learners' communicative ability, responses were scored using a holistic scale ranging from 0 to 3, accompanied by Can-Do style descriptors focusing on communicative appropriateness and speech intelligibility (i.e., word/utterance recognition), such as 'I can respond clearly and coherently to my partner(s) for most of the conversation (about 80% of the time)'. Participants received their task results in the form of rubrics upon completion of the session (see Appendix B).

Automated Scoring Model

The automated scores were generated using the Random Forest classifier originally developed and validated by Hayashi et al. (2024). The model was implemented using the Python package scikit-learn 1.1.3 (Pedregosa et al., 2011) and was trained to predict holistic human-rated DCT score categories (0–3). The original model used four types of speech features as predictors: token count, hesitation/filled pauses, syntactic complexity, and IBM Watson Speech-to-Text confidence scores—the degree of phonetic similarity to the acoustic model. During the original model development, 80% of the dataset was randomly allocated to model training and the remaining 20% to testing (Hayashi et al., 2024). The classifier used the default scikit-learn 1.1.3 settings for the number of trees (100), splitting criterion (Gini), maximum tree depth (None), minimum samples for splitting (2) and leaf nodes (1), bootstrap sampling (True), and class weighting (None). The exception was feature subsampling, with all features considered at each split rather than the default setting. No hyperparameter tuning or resampling for class imbalance was performed, nor was feature standardisation or scaling applied. Counts of hesitation/filled pauses and syntactic complexity features (e.g., xcomp) were expressed as rates per 100 words. In the present study, this previously trained model was applied to DCT1 and DCT2 without retraining or task-specific recalibration in order to examine its cross-task transferability (see Hayashi et al., 2024, for further details).

Learner Perception Questionnaire

A questionnaire modelled on Hayase et al. (2018) and Zhai and Ma (2022) was developed to elicit participants' perceptions of the automated scoring system and the feedback it provided. The instrument consisted of three main themes: (1) system, (2) task, and (3) feedback. These were further divided into seven subsections (Table 1 and Appendix C). All closed-ended items were rated on a 5-point Likert scale. In the pilot phase with 16 third-year undergraduate students aged 21–23, several participants expressed a desire to comment on aspects not covered by the questionnaire items. Accordingly, additional open-ended comment items were added to Cognitive Feedback (CF, 'Any additional information you would like to be included on the feedback page') and Perceived Usefulness (PU, 'Reasons for finding the test score and feedback useful'), respectively, to capture learners' perceptions more comprehensively. The open-ended responses were subsequently classified using the questionnaire framework shown in Table 1 as an a priori coding scheme. Comments were assigned to the category that best represented their primary focus, and the frequency of comments within each category was then calculated. For the open-ended PU item, recurring patterns in the responses were further identified inductively, resulting in six subcategories within the broader PU category. The open-ended responses and resulting classifications, including these six subcategories, were jointly reviewed and refined by the authors to ensure the appropriateness and consistency of the coding.

Table 1. Main themes and subsections of the questionnaire

Main theme	Subsection	Number of items
(1) System	Facilitating conditions (FC)	8
(2) Task	Task appropriateness (TA)	2
	Test-taking anxiety (TTA)	3
	Self-efficacy (SE)	2
(3) Feedback	Cognitive feedback (CF)	3
	Behavioural intention to use (BIU)	4
	Perceived usefulness (PU)	4

Note. See Appendix C for the full list of questionnaire items.

Procedure

Data were collected over two sessions (n = 59 per session). With the course lecturer’s permission and participants’ consent, the first author and two research assistants administered the speaking sessions in the classroom. Each session lasted approximately 30 minutes within a 90-minute class. Participants used tablets and headphones and were seated apart to minimise distractions. After 10–15 minutes of instructions and practice, they completed a warm-up task (e.g., name, age, and date of birth), followed by two DCTs (DCT1 and DCT2).

Results

For the DCT1, task-relevant keywords were grouped into semantically motivated categories corresponding to each prompt (e.g., action, destination, and social relations; transport, duration, and distance; and timing and duration), as shown in Table 2. These categories were used exclusively to characterise the content profiles of learner responses and were not incorporated into the automated scoring model. A parallel categorisation scheme was applied to DCT2, allowing direct comparison of content profiles across tasks. The categories included academic-related content (e.g., references to university admissions, studies, or examinations), affective or attitudinal expressions (e.g., emotional reactions to the situation), and reason- or explanation-related expressions used to justify or elaborate on those feelings.

Agreement between Human and Automated Scores

The random-forest scoring model developed by Hayashi et al. (2024) was used to predict the scores for both tasks (DCT1 and DCT2). Although only DCT1 was used in the original study, the features used to predict scores were not specific to its content. They included token counts, ratio of hesitation to total number of words, syntactic complexity (i.e., number of subordinate clauses), and confidence scores. Tables 3 and 4 present the confusion matrices for DCT1, and Tables 5 and 6 present those for DCT2, showing the distributions of the human and automated scores. Two raters independently scored the students’ responses to each task. Inter-rater agreement was high for both DCT1 and DCT2, with quadratically weighted kappa coefficients of .86 and .91, respectively. Exact agreement was 81.0% and 87.9%, while adjacent agreement (within one score band) was 100.0% and 99.1%, respectively. Gwet’s AC2, which is less sensitive to skewed score distributions, similarly indicated high agreement (DCT1 = .95; DCT2 = .98).

Table 2. Task-specific keyword categories and coverage for DCT1 and DCT2

Task	Prompt	Keyword category	Example lexical items (keywords)	Coverage (%)		
DCT1	Q1. A: <i>Do you have any suggestions for where to go?</i> B: _____ Q2. A: <i>Do you know how to get there?</i> B: _____ Q3. A: <i>When and how long is this holiday going to be?</i> B: _____	1-1. Action/activity	go, visit, go to	67.8		
		1-2. Destination (proper nouns/place nouns)	park, zoo, museum, school, sea, Hokkaido, hot spring, Fukuoka, Korea	23.3		
		1-3. Social relation	friend, you, we, together	48.6		
		2-1. Transport	by plane, train, bus, car, Shinkansen, express train	21.8		
		2-2. Duration (minutes/hours)	Take two hours, 30 minutes	25.3		
		2-3. Distance	far, near	1.4		
		3-1. Timing (month/period/season)	spring, summer, autumn, winter, break, holiday, weeks, weekend, month	27.6		
		3-2. Duration of period	for two weeks, months, days, years	35.6		
		DCT2	Q1. A: <i>I've got some good news! You know I've been waiting to hear from the university?</i> B: <i>Yes?</i> A: _____ Q2. B: <i>That means you're moving away. How do you feel about that?</i> A: _____	2-1. Academic/institutional content	university, offer, letter, study, enter, accepted, place, degree, internship, pass, exam	26.2
				Emotion/attitude adjectives	bad, sad, disappointed, happy, excited, looking forward to	46.5
Reason/explanation	because, so, reason, always wanted to, live alone			33.3		

Table 3. Confusion matrix between Rater 1 and automated scoring in DCT1

True score	Predicted 0	Predicted 1	Predicted 2	Predicted 3
0	1	0	1	2
1	0	1	4	11
2	0	1	12	25
3	0	0	10	48

Note: Rows represent the true scores assigned by human raters, and columns represent the scores predicted by the random forest model.

Table 4. Confusion matrix between Rater 2 and automated scoring in DCT1

True score	Predicted 0	Predicted 1	Predicted 2	Predicted 3
0	1	0	1	2
1	0	1	6	8
2	0	1	7	26
3	0	0	13	50

Table 5. Confusion matrix between Rater 1 and automated scoring in DCT2

True score	Predicted 0	Predicted 1	Predicted 2	Predicted 3
0	48	0	18	27
1	1	0	3	6
2	0	0	2	2
3	1	0	4	4

Table 6. Confusion matrix between Rater 2 and automated scoring in DCT2

True Score	Predicted 0	Predicted 1	Predicted 2	Predicted 3
0	48	0	18	23
1	1	0	1	8
2	0	0	4	5
3	1	0	4	3

In contrast, agreement between each human rater and the automated scores (hereafter, Human 1–Auto and Human 2–Auto) varied considerably depending on both the task and the agreement index used. For DCT1, quadratically weighted kappa coefficients were low for both Human 1–Auto ($\kappa = .21$) and Human 2–Auto ($\kappa = .23$). However, exact agreement was 53.4% and 50.9%, while adjacent agreement (within one score band) was 87.9% and 90.5%, respectively. Gwet's AC2 coefficients were substantially higher (.79 and .81, respectively) than the corresponding weighted kappa coefficients, suggesting that the low kappa values were at least partly influenced by the skewed score distributions.

The confusion matrices (Tables 3 and 4) showed that automated predictions were concentrated in the highest score category, with 86 of the 116 responses assigned a score of 3. The model correctly identified many responses rated as 3 by the human raters, but also frequently assigned a score of 3 to responses rated as 1 or 2, indicating a tendency towards the highest score category.

For DCT2, quadratically weighted kappa coefficients were again low for Human 1–Auto ($\kappa = .17$) and Human 2–Auto ($\kappa = .22$). However, unlike in DCT1, the complementary agreement indices were also substantially lower. Exact agreement was 46.6% and 47.4%, while adjacent agreement was 55.2% and 56.9%, respectively. Gwet's AC2 coefficients were also low (.26 and .29, respectively), indicating weak human–machine agreement in DCT2 even when using an agreement measure less affected by skewed score distributions. The confusion matrices (Tables 5 and 6) further revealed a discrepancy in the treatment of low-scoring responses. Human raters 1 and 2 assigned 93 and 89 of the 116 responses, respectively, to Score 0, whereas the automated model assigned 50 responses to Score 0. Although the automated model agreed with each human rater on 48 Score-0 responses, it assigned scores of 2 or 3 to many other responses receiving a human score of 0. For Human 1, for example, 18 of these responses were assigned an automated score of 2 and 27 an automated score of 3.

Inspection of the source transcripts indicated that the large number of zero scores in DCT2 did not primarily reflect an absence of analysable speech. Rather, because DCT performance was scored holistically across the multi-turn interaction, responses could receive a score of 0 when an initial turn was absent or failed to maintain discourse contingency with the preceding turn asking what the good news from the university was (e.g., “I can get good answer.”; “It is great you can make good memory.”; see Appendix A). In several such cases, learners nevertheless produced linguistically analysable speech in the subsequent turn. The automated model, in contrast, did not appear to be equally sensitive to this discourse-level relationship and could assign higher scores based on response-level linguistic and acoustic features. The fact that the automated model identified some human-rated Score-0 responses correctly but assigned Scores 2 or 3 to many others suggests that the discrepancy was not simply a general inability to identify low-performing responses. Rather, it may reflect differences between the discourse-level criteria applied in holistic human scoring and the response-level features represented in the automated model. This distinction also helps explain why lexical indices could be calculated for responses receiving a holistic human score of 0.

Task-Relevant Features associated with Human-Machine Agreement

To further explore the low agreement between human and automated scores, learner responses were examined for task-relevant linguistic and content features that may help explain variability in human–machine agreement across tasks. Specifically, lexical diversity (TTR: Type-Token Ratio) and content-related features (CWR: Content-Word Ratio; and KC: Keyword Coverage) were analysed. Task-relevant keywords were grouped into semantically motivated categories reflecting the communicative demands of each prompt or turn within each DCT (Table 2).

These features were used exclusively for descriptive content analysis and were not incorporated into the automated scoring model. The analysis examined the extent to which learner responses addressed task-relevant content and how this was reflected in human–machine discrepancies. For the correlational and regression analyses, lexical and semantic features were aggregated at the task level for each participant, consistent with the holistic task-level human and automated scores ($n = 116$ for each DCT).

As shown in Table 7, TTR and KC varied across turns in both tasks, with particularly high KC in DCT2 Turn 2. DCT1 responses frequently contained action/activity and social-relation vocabulary, whereas DCT2 elicited more emotion-, reason-, and academic/institution-related content (Table 2), indicating distinct task-relevant semantic profiles. Keyword categories also showed substantial variability across responses, supporting their use in the subsequent analyses.

Table 7. Means and standard deviations for the lexical and content-related features

		TTR	CWR	KC
DCT1	Turn 1	.74 (0.13)	.46 (0.08)	.24 (0.11)
	Turn 2	.81 (0.14)	.41 (0.12)	.28 (0.15)
	Turn 3	.81 (0.13)	.52 (0.11)	.42 (0.16)
DCT2	Turn 1	.85 (0.15)	.49 (0.15)	.21 (0.24)
	Turn 2	.79 (0.11)	.46 (0.09)	.48 (0.27)

Note. TTR: Type-Token Ratio; CWR: Content-Word Ratio; Keyword Coverage. Values are presented as means (standard deviations).

Given the high human-rater agreement in both DCTs, the mean of the two raters’ task-level scores was used as the human score in the correlational and regression analyses. Participant-level Pearson correlations (Table 8) showed little association between CWR and either score type, whereas TTR was negatively associated with automated scores in both tasks. For DCT1, action-related expressions and social references were positively associated with both human and automated scores, whereas duration-time and transport-related expressions were associated only with human scores. For DCT2, academic-related vocabulary was positively associated with both score types, whereas emotion- and reason-related categories showed little association with either.

To further examine whether semantic variables contributed to score prediction beyond lexical features, hierarchical multiple regression analyses were conducted separately for human and automated scores within each task. CWR and TTR were entered at Step 1, followed by task-specific semantic variables at Step 2 (Tables 9–12).

For DCT1, adding semantic variables significantly improved the prediction of human scores, accounting for an additional 26% of the variance (Table 9). Action-related expressions, social references, duration-time expressions, and transport-related expressions were significant positive predictors. Semantic variables also explained an additional 15% of the variance in automated scores (Table 10), with CWR, action-related expressions, and social references as significant positive predictors and TTR as a significant negative predictor. Thus, automated scores reflected some of the same task-relevant semantic features as human scores, although fewer semantic predictors made significant unique contributions.

Table 8. Correlations between human and automated scores and lexical and semantic features by task

Features			DCT1 (n = 116)		DCT2 (n = 116)	
			human scores	auto scores	human scores	auto scores
Content-word ratio			.00	.04	.15	.05
Type-token ratio			−.15	−.35***	−.21*	−.37***
DCT1	Turn 1	action	.31***	.32***		
		destination	.09	.13		
		social relation	.26**	.30***		
		distance	.07	.04		
	Turn 2	duration_time	.22*	−.09	NA	
		transport	.30**	.16		
	Turn 3	duration_period	.10	.02		
		timing	.16	.10		
DCT2	Turn 1	academic			.41***	.38***
	Turn 2	emotion	NA		.15	.10
		reason			.13	.09

p* < .05. *p* < .01. ****p* < .001.

Table 9. Hierarchical regression results for human scores on DCT 1

Variable	<i>B</i>	95% CI for <i>B</i>		SE <i>B</i>	β	R^2	ΔR^2
		LL	UL				
Step 1						.03	.03
Constant	2.87***	1.49	4.25	.70			
Content-word ratio	.64	−2.10	3.37	1.38	.04		
TTR	−1.42	−3.06	.23	.83	−.16		
Step 2						.28	.26***
Constant	.16	−1.39	1.70	.78			
Content-word ratio	1.81	−.72	4.35	1.28	.13		
TTR	.18	−1.43	1.79	.81	.02		
Turn 1_action	.28***	.12	.43	.08	.31		
Turn 1_social relation	.19**	.06	.32	.06	.26		
Turn 2_duration_time	.16*	.04	.28	.06	.21		
Turn 2_transport	.18**	.06	.29	.06	.25		

Note. *B* = unstandardized regression coefficient; SE *B* = standard error; β = standardized regression coefficient.

p* < .05. *p* < .01. ****p* < .001.

Table 10. Hierarchical regression results for automated scores on DCT 1

Variable	<i>B</i>	95% CI for <i>B</i>		SE <i>B</i>	β	<i>R</i> ²	ΔR^2
		LL	UL				
Step 1						.14	
Constant	3.40***	2.52	4.27	.44			
Content-word ratio	1.39	−.35	3.12	.88	.14		
TTR	−2.24***	−3.28	−1.19	.53	−.39		
Step 2						.29	.15***
Constant	1.99***	.94	3.03	.53			
Content-word ratio	2.36**	.65	4.07	.86	.24		
TTR	−1.46**	−2.54	−.38	.55	−.25		
Turn 1_action	.12*	.01	.22	.05	.19		
Turn 1_social relation	.16***	.07	.24	.04	.32		
Turn 2_duration_time	−.05	−.13	.03	.04	−.10		
Turn 2_transport	.06	−.02	.13	.04	.12		

p* < .05. *p* < .01. ****p* < .001.

Table 11. Hierarchical regression results for human scores on DCT 2

Variable	<i>B</i>	95% CI for <i>B</i>		SE <i>B</i>	β	<i>R</i> ²	ΔR^2
		LL	UL				
Step 1						.07	.07*
Constant	.78	−.62	2.18	.71			
Content word ratio	1.93	−.20	4.07	1.08	.16		
TTR	−1.70*	−3.09	−.32	.70	−.22		
Step 2						.20	.13***
Constant	.13	−1.22	1.47	.68			
Content word ratio	1.50	−.51	3.50	1.01	.13		
TTR	−.89	−2.24	.46	.68	−.12		
Turn 1_academic	.41***	.21	.60	.10	.37		

p* < .05. *p* < .01. ****p* < .001.

For DCT2, academic-related vocabulary significantly improved the prediction of both human and automated scores, accounting for an additional 13% and 8% of the variance, respectively (Tables 11 and 12). Academic-related vocabulary was a significant positive predictor in both models, while TTR remained a significant negative predictor of automated scores. Overall, task-relevant semantic content contributed significantly to both human and automated scores in DCT2, while automated scores also showed a negative association with lexical diversity.

Table 12. Hierarchical regression results for automated scores on DCT 2

Variable	<i>B</i>	95% CI for <i>B</i>		SE <i>B</i>	β	R^2	ΔR^2
		LL	UL				
Step 1						.14	
Constant	4.14***	2.07	6.21	1.04			
Content-word ratio	1.21	−1.95	4.36	1.59	.07		
TTR	−4.37***	−6.42	−2.32	1.03	−.37		
Step 2						.22	.08**
Constant	3.36**	1.32	5.39	1.03			
Content-word ratio	.68	−2.36	3.72	1.53	.04		
TTR	−3.38**	−5.43	−1.33	1.03	−.29		
Turn 1_academic	.49**	.20	.78	.15	.29		

* $p < .05$. ** $p < .01$. *** $p < .001$.

Because the human scores were derived from ordinal ratings and the original DCT2 ratings showed substantial floor concentration, proportional-odds ordinal logistic regression analyses were additionally conducted as a sensitivity check using each rater’s original 0–3 score separately. The sensitivity analyses yielded substantively consistent results: all four significant semantic predictors in DCT1 remained significant positive predictors for both raters, as did academic-related vocabulary in DCT2.¹ These results indicate that the principal semantic patterns identified for the human ratings were robust to the ordinal nature and floor concentration of the underlying ratings. The hierarchical linear regression analyses were therefore retained as the primary analyses for comparing the relative contributions of lexical and semantic predictors to human and automated scores.

Learner Perceptions

The results of the learner perception questionnaire were analysed descriptively (Table 13). In five of the seven categories (FC, TA, SE, CF, and PU), the participants generally expressed positive perceptions, with mean ratings around 4.0 or higher. Exceptions were FC3 (*I was able to take the test without being disturbed by surrounding noise*) and TA2 (*I was able to understand what content I was expected to respond with*), both of which received lower evaluations. In the TTA category, participants’ ratings clustered around the midpoint, indicating ambivalent attitudes towards the anxiety associated with performing an automated speaking task of this nature. In the BIU category, the ratings were lower for BIU2 (*I think I will use automated-scoring speaking tasks such as these frequently in the future*) and BIU3 (*I intend*

¹Odds ratios (ORs) and 95% confidence intervals (CIs) were as follows. In DCT1, action-related expressions (Rater 1: OR = 2.11, 95% CI [1.34, 3.31], $p = .001$; Rater 2: OR = 1.97, 95% CI [1.23, 3.15], $p = .005$), social references (Rater 1: OR = 1.70, 95% CI [1.16, 2.48], $p = .006$; Rater 2: OR = 1.65, 95% CI [1.13, 2.41], $p = .009$), duration-time expressions (Rater 1: OR = 1.47, 95% CI [1.02, 2.10], $p = .038$; Rater 2: OR = 1.78, 95% CI [1.20, 2.62], $p = .004$), and transport-related vocabulary (Rater 1: OR = 1.76, 95% CI [1.23, 2.52], $p = .002$; Rater 2: OR = 1.47, 95% CI [1.03, 2.09], $p = .032$) were significant positive predictors. In DCT2, academic-related vocabulary was significant for both Rater 1 (OR = 2.97, 95% CI [1.65, 5.37], $p < .001$) and Rater 2 (OR = 2.29, 95% CI [1.38, 3.80], $p = .001$).

Table 13. Descriptive statistics for the questionnaire items

Group	No.	Mean (SD)	Raw counts of responses (n = 118)				
			5	4	3	2	1
Facilitating Conditions (FC)	FC1	4.73 (0.61)	94	18	4	2	0
	FC2	4.85 (0.45)	104	10	4	0	0
	FC3	3.31 (1.04)	15	38	38	23	4
	FC4	4.85 (0.43)	103	12	3	0	0
	FC5	4.75 (0.51)	93	21	4	0	0
	FC6	4.74 (0.56)	93	20	4	1	0
	FC7	4.56 (0.88)	87	18	7	4	2
	FC8	4.41 (0.86)	71	29	14	3	1
Task Appropriateness (TA)	TA1	4.08 (0.77)	39	51	27	1	0
	TA2	3.61 (0.98)	23	43	37	13	2
Test-taking Anxiety (TTA)	TTA1	2.49 (1.08)	4	21	25	47	21
	TTA2	2.67 (1.16)	7	24	30	37	20
	TTA3	2.46 (1.13)	4	20	30	36	28
Self-efficacy (SE)	SE1	4.55 (0.70)	76	34	5	3	0
	SE2	4.13 (1.00)	54	36	19	7	2
Cognitive Feedback (CF)	CF1	3.62 (1.01)	24	45	31	16	2
	CF2	4.07 (0.97)	49	37	24	7	1
Behavioral Intention to Use (BIU)	BIU1	3.61 (0.94)	23	39	45	9	2
	BIU2	3.37 (0.91)	15	32	55	14	2
	BIU3	3.04 (0.89)	7	23	61	22	5
	BIU4	3.96 (0.91)	37	47	27	6	1
Perceived Usefulness (PU)	PU1	3.90 (1.03)	40	41	24	11	2
	PU2	4.03 (0.93)	43	45	22	7	1
	PU3	4.18 (0.85)	52	38	25	3	0

*5: strongly agree; 4: agree; 3: neither agree nor disagree; 2: disagree; 1: strongly disagree.

to continue using automated-scoring speaking tasks such as these in the future), suggesting uncertainty about their utility in autonomous learning contexts.

Analysis of the optional open-ended comments submitted by 18 participants on the feedback page revealed several suggestions for additional information they wished to receive (CF3). Seven participants requested model responses to help them understand target performance levels and compare their answers with exemplary ones, while six requested transcripts of their

spoken responses for review and reflection. Two participants requested learning advice, such as appropriate expressions or guidance on improving speaking ability, two requested pronunciation feedback, and one suggested linking the feedback to established proficiency tests such as EIKEN.

Analysis of another optional open-ended item (PU4) revealed six recurring themes regarding participants' perceptions of feedback (Appendix D). Eighteen participants valued the opportunity to understand their current English proficiency and the immediacy of the results, as illustrated by the following comments:

'I understood my current level of English ability and received immediate results.'

'The test allowed me to view my English performance objectively.'

Twenty-two participants highlighted the diagnostic and specific nature of the feedback, particularly its identification of strengths and weaknesses and suggestions for improvement:

'The feedback clearly showed what I did well and what I still needed to improve.'

'The feedback accurately identified the parts I lacked confidence in.'

Fourteen participants reported that the feedback encouraged reflection and informed future learning (e.g., 'Looking at the scoring criteria influenced how I plan to study and my overall learning attitude'). Eleven valued the clarity, transparency, and perceived fairness of the scoring process (e.g., 'The feedback clearly explained the rationale behind each score'), nine emphasised its practical usefulness (e.g., 'The comments were specific enough to use as reference material'), and seven reported increased motivation and confidence (e.g., 'The feedback encouraged me to continue improving'). A complete list of comments is provided in Appendix D.

Discussion

The present study examined the cross-task transferability of an automated scoring model developed for a multi-turn DCT (Hayashi et al., 2024). Across agreement indices, human-machine agreement was considerably stronger in DCT1 than in DCT2, whereas human-human agreement remained high across both tasks. This pattern indicates reduced model alignment with human judgments when transferred to DCT2 without retraining, underscoring the importance of cross-task validation in automated speaking assessment.

Limitations of the Current Automated Scoring Model in DCT-based Assessment

Unlike human raters, the automated model did not appear to adjust its scoring behaviour adequately to all aspects of the communicative demands of the two tasks. This finding aligns with previous research highlighting the task-sensitive nature of holistic human judgments, in which raters integrate multiple cues, including content relevance and communicative appropriateness, when forming overall evaluations (Kondo & Ishii, 2017; Sawaki & Koizumi, 2017). The transcript inspection provided a particularly clear illustration of this mismatch in DCT2. Importantly, sensitivity to discourse continuity was inherent in the original human-scoring construct rather than being introduced as a post-hoc interpretation in the present study. In the original model-development study (Hayashi et al., 2024), holistic scores reflected learners' ability not only to produce intelligible and coherent responses but also to sustain and advance the discourse across turns. A breakdown at an earlier turn could therefore result in a score of 0 even

when a learner subsequently produced linguistically analysable speech. The automated model, however, did not appear to be equally sensitive to this discourse-level relationship and could assign higher scores based on response-level linguistic and acoustic features. Nevertheless, the automated model did identify some human-rated Score-0 responses correctly but assigned Scores 2 or 3 to many others. The discrepancy observed in DCT2 may therefore reflect differential sensitivity to the discourse-level construct represented in the human scoring criteria, rather than a general inability of the model to identify low-performing responses.

The correlational and hierarchical regression analyses provided a more nuanced account of the relationship between task-relevant semantic content and scoring. For DCT1, action-related expressions and social references were associated with both human and automated scores, whereas duration-time expressions and transport-related vocabulary contributed more specifically to human scores. Semantic variables also explained more additional variance in human than automated scores. In DCT2, however, academic-related vocabulary was associated with both score types and significantly improved the prediction of both human and automated scores. Thus, the automated model was not uniformly insensitive to task-relevant semantic content. Rather, its sensitivity varied across tasks and semantic features, suggesting that the central limitation extends beyond semantic sensitivity alone.

In this respect, the observed decline in human-machine agreement across tasks could be better understood in light of the model-transfer conditions adopted in the present study. The existing scoring model was intentionally applied to the new DCT without task-specific retraining or recalibration in order to examine its transferability in its existing form, as specified in the research questions. Consequently, the current findings cannot determine whether the reduced agreement reflects limitations in the existing predictor set or Random Forest modelling approach, or whether performance could be improved through task-specific adaptation to DCT2. Although beyond the scope of the current study, retraining the model on DCT2 responses might, for example, have enabled it to learn task-specific relationships and interactions among the existing predictors. The present findings therefore demonstrate limited cross-task transferability without task-specific adaptation, rather than establishing that the underlying predictors or modelling approach are inherently unsuitable for reliable scoring across DCTs.

The findings, taken together, raise questions concerning the validity of the automated DCT assessment employed in the present study, but they also refine the nature of the potential construct underrepresentation. The results do not suggest that the automated model was simply insensitive to task-relevant semantic information: the analyses of both DCTs demonstrate that at least some task-relevant semantic features were meaningfully reflected in automated as well as human scores. Rather, from a construct-representation perspective, the more fundamental limitation may concern the model's ability to represent how linguistic and semantic content functions within a multi-turn discourse context. The target construct involves not only producing relevant linguistic material but also fulfilling task-specific communicative demands and maintaining discourse contingency across turns. The DCT2 transcript inspection illustrates this distinction particularly clearly: linguistically analysable and even task-related material could contribute to an automated score even when a response failed to function appropriately in relation to the preceding turn and was consequently assigned a low holistic score by human raters. This mismatch is particularly relevant because discourse continuity formed part of the holistic human-scoring construct used in the original model-development study, whereas it was not directly represented among the linguistic and acoustic predictors available to the automated model (Hayashi et al., 2024). Interaction elicited in such contexts bears some resemblance to naturally occurring discourse in allowing a degree of spontaneity and unpredictability in responses (Hayashi et al., 2024; Ogiermann, 2018). This greater response variability poses

particular challenges for maintaining rating quality and supporting valid score interpretations, an issue also documented in research on the automated assessment of spontaneous speech (Norris & Davis, 2024; Zechner & Loukina, 2020).

One possible direction for addressing this aspect of construct underrepresentation can be found in recent advances in automated scoring research that model content relevance and task alignment more explicitly. For example, Lu et al. (2025) introduced a Multifaceted Relevance Model that incorporates task-specific response segmentation and relevance estimation between learner responses and exemplar answers or visual prompts. Such models explicitly operationalise the semantic alignment between task prompts and responses, going beyond surface-level linguistic or acoustic features. For multi-turn DCT assessment, however, the present findings suggest that modelling semantic relevance alone may not be sufficient: automated scoring may also need to represent the relationship between a response and the preceding discourse in order to evaluate whether task-relevant content functions communicatively within the interaction. Although advanced architectures offer promising directions for improving such sensitivity, they also require substantial task-specific resources, such as curated exemplar responses or multi-modal inputs, which were not available in the present study.

These findings thus demonstrate the limitations of applying the current automated scoring model (Hayashi et al., 2024) across different DCT-based speaking tasks. The multiple agreement indices showed that the low weighted kappa coefficients did not tell the same story across the two tasks: agreement in DCT1 was considerably stronger when assessed using indices less affected by skewed score distributions, whereas multiple indices converged in showing substantially weaker human–machine agreement in DCT2. The regression analyses suggest that differences between human and automated scoring cannot be explained simply in terms of sensitivity to task-relevant semantic content versus reliance on surface-level features. In DCT2, academic-related content likewise contributed significantly to both. Nevertheless, the sharp human–machine disagreement in DCT2, together with the transcript evidence, suggests that the automated model did not adequately represent how otherwise analysable linguistic and semantic content functioned within the multi-turn discourse. More broadly, these findings highlight the importance of construct alignment in cross-task automated assessment: transferability requires not only stable agreement between automated and human scores across tasks but also adequate representation of the construct-relevant relationships among linguistic content, task fulfilment, and discourse context. The identified limitations in scoring validity, however, do not necessarily predict how learners experience or value a system. Therefore, the following section focuses on learners’ perceptions of the automated speaking test and its feedback.

Learner Perceptions of the Automated Speaking Test and its Feedback

The findings of the present study align with those of previous research (e.g., Gu et al., 2021; Sawaki & Koizumi, 2017; Zhai & Ma, 2022), indicating that most participants perceived automated feedback on their performance in multi-turn DCTs as beneficial. Nevertheless, several participants reported anxiety and discomfort associated with the testing conditions, such as background noise, and with the automated, time-regulated nature of the task, which they perhaps perceived as pressurised, potentially limiting its usefulness for sustained autonomous learning. These concerns are distinctive to the current study and likely stem from the testing arrangement, as the speaking session was conducted in a large classroom with approximately 60 participants simultaneously completing identical tasks. As suggested by the SE items, such issues may be mitigated if learners undertake the speaking task individually in a quiet environment, which, in turn, fosters more positive attitudes towards the continued use of the task and its feedback.

Participants also assigned relatively low ratings to the feedback's perceived utility for self-directed learning, a pattern consistent with previous studies in which learners' motivation to engage with feedback declined when score reporting was delayed (e.g., Sawaki & Koizumi, 2017) or when feedback was regarded as overly general (Ranalli, 2021; Zhang, 2020). In contrast, the feedback in the present study was generated immediately after task completion and included both task-specific and task-general descriptors, as well as an overall score (Appendix B). However, what may have been lacking was more explicit guidance for learning embedded within the feedback. Nonetheless, several participants noted that they could identify areas for improvement based on their feedback. This was reflected in optional comments, in which participants reported that feedback was most useful when their perceived strengths and weaknesses aligned with the scoring criteria. Therefore, a longitudinal investigation could examine how learners subsequently utilise such feedback in their ongoing studies and skill development, and whether such use results in beneficial washback.

Another notable finding was that participants valued feedback as an indicator of their current level of communicative ability. One interpretation of this appreciation is that participants extrapolated their level of communicative ability from the scoring rubrics presented as feedback, which they perceived as trustworthy, transparent, and fair. While there has been increasing research interest in AI-based feedback in speaking assessment, these observations suggest that rule-based feedback in the form of rubrics can assist learners by providing interpretable, pedagogically transparent feedback (Hayashi, 2025; Vieriu & Petrea, 2025). At the same time, the discrepancy between learners' positive perceptions of the feedback and the relatively weak human-machine agreement, particularly for DCT2, raises an important validity concern regarding the potential consequences of score interpretation and use (Messick, 1995). In this context, perceived transparency and credibility should not be taken as evidence that the underlying automated scores adequately represent the intended construct. Indeed, feedback that learners regard as credible may influence subsequent learning even when the scores on which it is based diverge from expert judgments. The positive learner perceptions observed here should therefore be interpreted cautiously: while they support the usability and perceived pedagogical value of the system, they also highlight the need to ensure sufficient scoring validity before automated feedback is used to guide learners' autonomous practice.

However, the feedback system used in the present study has several limitations. As noted in participants' comments, transparency could be improved by providing model responses and transcripts alongside the rubrics, which may help learners notice gaps between their production and target performance. Some participants also expressed a preference for criterion-referenced scores, particularly for pronunciation. Although the system provides feedback on intelligibility, the link between intelligibility and pronunciation may not have been sufficiently explicit. One possible improvement would be to incorporate content and intelligibility as separate criteria within the holistic rubrics. In addition, aligning the scoring model with established proficiency frameworks (e.g., the EIKEN test) may enhance perceived reliability. Finally, as the system is currently available only on tablets, limited accessibility may have reduced learners' willingness to engage with the tasks. A web-based version is therefore under development to support broader and more flexible use. More fundamentally, improvements to feedback presentation should be accompanied by improvements in scoring validity, particularly if learners are encouraged to use the feedback for autonomous learning.

Conclusion

Evaluation of the current automated scoring model beyond single-task validation has shown that its limitations cannot be attributed simply to an inability to capture task-relevant semantic

information, as automated scores showed sensitivity to particular task-relevant semantic features in both DCTs. Rather, the findings point to a broader construct-alignment challenge: representing not only relevant linguistic and semantic content but also how that content functions within a multi-turn interaction. In particular, the contrast between human and automated scoring suggests the importance of capturing discourse contingency when evaluating task-specific communicative performance. Building on this limitation, ongoing development of the system is exploring a branching multi-turn design in a web-based environment, in which automated scoring of responses triggers response-specific prompts in subsequent turns.

Although learners generally perceived the system positively, such perceptions should not be taken as evidence of scoring validity. Nevertheless, its rule-based framework provided relatively transparent and interpretable feedback. Such systems may therefore be better positioned as learning-support tools than as fully reliable assessment instruments. Future research should examine whether task-specific adaptation and greater representation of discourse-level relationships can improve scoring validity while retaining the immediacy and transparency that make automated feedback potentially valuable for autonomous learning.

Acknowledgements

The authors express their sincere gratitude to Akira Murakami, Jonathan Moxon, Nanako Hayase, Ryota Sao, Kenshiro Tokushima, and Mikikazu Uchida for their cooperation and support throughout the study; and to all participants for their hard work throughout the study. Finally, we thank the Admission Center of Saga University for administrative and technical support, the Research and Development Center of the Saga Densen Center (<https://www.sdcns.co.jp/>) for their expertise in systems engineering, and Editage (www.editage.com) for English language editing.

Declaration of Use of Generative AI

A generative AI tool (ChatGPT) was used for textual editing only in order to support language editing and improve clarity. A professional proofreading service (Editage) was also used to refine the manuscript.

References

- Blum-Kulka, S. (1982). Learning to say what you mean in a second language: A study of the speech act performance of learners of Hebrew as a second language. *Applied Linguistics*, 3, 29–59. <https://doi.org/10.1093/applin/III.1.29>
- Blum-Kulka, S., & Olshtain, E. (1984). Requests and apologies: A cross-cultural study of speech act realization patterns (CCSARP). *Applied Linguistics*, 5, 196–213. <https://doi.org/10.1093/applin/5.3.196>
- Chapelle, C. A. (2012). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 21–33). Routledge.
- Chen, L., Zechner, K., Yoon, S. Y., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C. M., & Ma, M. (2018). Automated scoring of nonnative speech using the *SpeechRater*SM v. 5.0 engine. *ETS Research Report Series*, 2018(1), 1–31. <https://doi.org/10.1002/ets2.12198>
- Council of Europe. (2020). *The common European framework of reference for languages: Learning, teaching, assessment—Companion volume*.
- Cucchiarini, C., Neri, A., & Strik, H. (2009). Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback. *Speech Communication*, 51, 853–863. <https://doi.org/10.1016/j.specom.2009.03.003>
- Czyż, K., Derkacz, M., Smółka, J., Łukasik, E., & Skublewska-Paszkowska, M. (2019). Speech recognition APIs in the context of using English as a second language. In Z. Czyż & M. Badurowicz (Eds.), *Computational methods in engineering science* (pp. 29–40). Politechnika Lubelska.

- Ding, D., & Yusof, M. B. (2025). Investigating the role of AI-powered conversation bots in enhancing L2 speaking skills and reducing speaking anxiety: A mixed methods study. *Humanities and Social Sciences Communications*, 12(1), Article 1223. <https://doi.org/10.1057/s41599-025-05550-z>
- Eslami, Z. R., & Mirzaei, A. (2014). Speech act data collection in a non-Western context: Oral and written DCTs in the Persian language. *International Journal of Language Testing*, 4, 137–154.
- ETS. (2025). *Mapping the TOEIC® tests on the CEFR*. <https://www.ets.org/pdfs/toEIC/toEIC-mapping-cefr-reference.pdf>
- Golato, A. (2003). Studying compliment responses: A comparison of DCTs and recordings of naturally occurring talk. *Applied Linguistics*, 24, 90–121. <https://doi.org/10.1093/applin/24.1.90>
- Grainger, K., Kerkam, Z., Mansor, F., & Mills, S. (2015). Offering and hospitality in Arabic and English. *Journal of Politeness Research*, 11, 41–70. <https://doi.org/10.1515/pr-2015-0003>
- Gu, L., Davis, L., Tao, J., & Zechner, K. (2021). Using spoken language technology for generating feedback to prepare for the TOEFL iBT® test: A user perception study. *Assessment in Education: Principles, Policy & Practice*, 28, 58–76. <https://doi.org/10.1080/0969594X.2020.1735995>
- Hayase, H., Hayashi, Y., & Eguchi, M. (2018). Yonginou wo tou eigo CBT nyushikaihatsu ni muketa torikumi [Towards the Implementation of a Computerized Integrated English Language Entrance Examination]. *LET Kyushu-Okinawa BULLETIN* 18, 15–29. https://doi.org/10.24716/letko.18.0_15
- Hayashi, Y. (2019). Investigating effects of working memory training on foreign language development. *The Modern Language Journal*, 103, 665–685. <https://doi.org/10.1111/modl.12584>
- Hayashi, Y. (2025). SLA no shiten kara kangaeru seisei AI no kyōikuteki katsuyō to sono ryūiten [Educational use of generative AI from an SLA perspective and points to consider]. *Eigo Kyoiku*, 74(6), 62–64.
- Hayashi, Y., Kondo, Y., & Ishii, Y. (2024). Automated speech scoring of dialogue response by Japanese learners of English as a foreign language. *Innovation in Language Learning and Teaching*, 18, 32–46. <https://doi.org/10.1080/17501229.2023.2217181>
- Houck, N., & Gass, S. M. (1996). Non-native refusals: A methodological perspective. In P. Jordens (Ed.), *Speech acts across cultures: Challenges to communication in a second language* (pp. 45–64). Mouton de Gruyter.
- Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to LanguEdge assessment. *Language Testing*, 26, 31–73. <https://doi.org/10.1177/0265532208097336>
- Jiang, M. Y.-C., Jong, M. S.-Y., Wu, N., Shen, B., Chai, C.-S., Lau, W. W.-F., & Huang, B. (2022). Integrating automatic speech recognition technology into vocabulary learning in a flipped English class for Chinese college students. *Frontiers in Psychology*, 13, Article 902429. <https://doi.org/10.3389/fpsyg.2022.902429>
- Kasper, G. (2000). Data collection in pragmatics research. In H. Spencer-Oatey (Ed.), *Culturally speaking: Managing rapport through talk across cultures* (pp. 316–341). Continuum.
- Knoch, U., Fairbairn, J., & Jin, Y. (2021). *Scoring second language spoken and written performance: Issues, options and directions*. Equinox.
- Kondo, Y., & Ishii, Y. (2017). Eigo gakushusha no hatsuwa jidousaiten shisutemu no kaishatsu to eigokyoiku puroguramu eno dounyu no kentou [Presenting practicality of automated speech scoring system in English language education program]. *Language Education & Technology*, 54, 23–40.
- Liu, Y., Binti Ab Rahman, F., & Binti Mohamad Zain, F. (2025). A systematic literature review of research on automatic speech recognition in EFL pronunciation. *Cogent Education*, 12, Article 2466288. <https://doi.org/10.1080/2331186X.2025.2466288>
- Lu, H.-C., Lin, J.-K., Lin, H.-Y., Wang, C.-C., & Chen, B. (2025). Advancing automated speaking assessment leveraging multifaceted relevance and grammar information. *arXiv preprint arXiv:2506.16285*. <https://doi.org/10.48550/arXiv.2506.16285>
- Macháček, D., Dabre, R., & Bojar, O. (2023). Turning Whisper into a real-time transcription system. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 17–24).
- McGuire, M., & Larson-Hall, J. (2025). Assessing Whisper automatic speech recognition and WER scoring for elicited imitation: Steps toward automation. *Research Methods in Applied Linguistics*, 4, Article 100197. <https://doi.org/10.1016/j.rmal.2025.100197>

- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Mingyan, M., Noordin, N., & Razali, A. B. (2025). Improving EFL speaking performance among undergraduate students with an AI-powered mobile app in after-class assignments: An empirical investigation. *Humanities and Social Sciences Communications*, 12, 1–15. <https://doi.org/10.1057/s41599-025-04688-0>
- Ngadiso, Wijaya, A. S., Marduma, A., Nisa, A. K., & Fahmi, A. N. (2023). The implementation of automated speech recognition (ASR) in ELT classrooms: A systematic literature review from 2012–2023. *JETAL: Journal of English Teaching and Applied Linguistics*, 5, 41–56. <https://doi.org/10.36663/jetal.v5i1.609>
- Norris, J. M., & Davis, L. (2024). Assessing second language speaking at ETS. In J. M. Norris & L. Davis (Eds.), *Challenges and innovations in speaking assessment* (pp. 1–18). Routledge.
- Ogiermann, E. (2018). Discourse completion tasks. In A. H. Jucker, K. P. Schneider, & W. Bublitz (Eds.), *Methods in pragmatics* (pp. 229–255). De Gruyter Mouton.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 28492–28518).
- Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing*, 52, Article 100816. <https://doi.org/10.1016/j.jslw.2021.100816>
- Rodrigues, L., Xavier, C., Costa, N., Gašević, D., & Mello, R. F. (2025). Is GPT-4 fair? An empirical analysis in automatic short answer grading. *Computers and Education: Artificial Intelligence*, 8, Article 100428. <https://doi.org/10.1016/j.caeai.2025.100428>
- Sawaki, Y., & Koizumi, R. (2017). Providing test performance feedback that bridges assessment and instruction: The case of two standardized English language tests in Japan. *Language Assessment Quarterly*, 14, 234–256. <https://doi.org/10.1080/15434303.2017.1348504>
- Sweeney, E., & Hua, Z. (2010). Accommodating toward your audience: Do native speakers of English know how to accommodate their communication strategies toward nonnative speakers of English? *The Journal of Business Communication*, 47, 477–504. <https://doi.org/10.1177/0021943610377308>
- Sweeney, E., & Hua, Z. (2016). Discourse completion tasks. In Z. Hua (Ed.), *Research methods in intercultural communication: A practical guide* (pp. 212–222). Wiley Blackwell.
- Vieriu, A. M., & Petrea, G. (2025). The impact of artificial intelligence (AI) on students' academic development. *Education Sciences*, 15, Article 343. <https://doi.org/10.3390/educsci15030343>
- Yuan, Y. (2001). An inquiry into empirical pragmatics data-gathering methods: Written DCTs, oral DCTs, field notes, and natural conversations. *Journal of Pragmatics*, 33, 271–292. [https://doi.org/10.1016/S0378-2166\(00\)00031-X](https://doi.org/10.1016/S0378-2166(00)00031-X)
- Zechner, K., & Hsieh, C. N. (2024). Automated scoring and feedback for spoken language. In *The Routledge international handbook of automated essay evaluation* (pp. 141–160). Routledge.
- Zechner, K., & Loukina, A. (2020). Automated scoring of extended spontaneous speech. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring* (pp. 365–382). Chapman and Hall/CRC.
- Zhai, N., & Ma, X. (2022). Automated writing evaluation (AWE) feedback: A systematic investigation of college students' acceptance. *Computer Assisted Language Learning*, 35, 2817–2842. <https://doi.org/10.1080/09588221.2021.1897019>
- Zhang, Z. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing*, 43, Article 100439. <https://doi.org/10.1016/j.asw.2019.100439>

Author Bio

Yuko Hayashi is an associate professor in applied linguistics at the Faculty of Education, Saga University, Japan. Since completing her doctoral studies at the University of Oxford in 2012, she has conducted research on working memory training and automated scoring of speech. Her teaching focuses on foreign language teaching methods and research methods. yhayashi@cc.saga-u.ac.jp.

Yusuke Kondo is an associate professor at Waseda University. He received his BA, MEd, and PhD from the School of Education, Waseda University. His research interests include language testing, automated scoring of language learners' performance, and learner corpora. yusukekondo@waseda.jp.

Yutaka Ishii is an associate professor at the Graduate School of Education, Chiba University, Japan. His research interests lie in English language education and educational technology. yishii@chiba-u.jp.

Appendix A. DCTs used in the Present Study

DCT1

- *Instructions: You (B) are making plans for an upcoming holiday with A. Please provide an appropriate response within 30 seconds.*

Turn 1 A: Do you have any suggestions for where to go?

B:  _____

Turn 2 A: Do you know how to get there?

B:  _____

Turn 3 A: When and how long is this holiday going to be?

B:  _____

DCT2

- *Instructions: You (A) are having a conversation with B. Please provide an appropriate response within 30 seconds.*

A: I've got some good news! You know I've been waiting to hear from the university?

B: Yes?

Turn 1 A:  _____

B: That means you're moving away. How do you feel about that?

Turn 2 A:  _____

Appendix B. Scoring Rubrics for DCT1 and DCT2

DCT1

3 points	• Interacts by responding intelligibly and coherently to the other interlocutor(s) for at least 80 percent of the time. • Responds to the first question by either naming an actual destination and/or providing a list of candidates alone or together with expansions in order to sustain and advance the discourse, such as reasons for, and/or access to the said destination. • Follows on from Turn 1 by providing information about both access to the said destination(s) and suggesting when and how long the suggested holiday will take place.
2 points	• Interacts by responding intelligibly and coherently to interlocutor(s) for at least 50 percent of the time but exhibits some noticeable lapses in the expression of ideas. • Responds to the first question with an actual destination and/or a list of candidates, followed by providing information about access to the said destination(s) only. • Follows on from Turn 1 by providing information about both access to the destination(s) (Turn 2) and the timing and duration of the suggested holiday (Turn 3) in a short, discrete utterance consisting of erroneous structures or mispronunciations impeding intelligibility/comprehensibility.
1 point	• Interacts by responding intelligibly and coherently to interlocutor(s) for at least 30 percent of the time. • Responds to the first question, with an actual destination and/or a list of candidates, followed by silence or lack of intelligible or coherent responses at the subsequent turns. • Is unable to continue from Turn 1 by either providing information about access to the said destination(s) or the timing and duration of the suggested holiday.
0 points	• Makes no attempt to respond or the response is unintelligible or unrelated to the topic. • Makes no or little attempt to sustain or advance the discourse at Turn 1 and thereafter, falling silent and/or providing incomplete responses.

DCT2

3 points	• Interacts by responding intelligibly and coherently to the other interlocutor(s) for at least 80% of the time. • Responds by revealing the nature of good news (Turn 1) and expressing how s/he feels about moving away with expansions (e.g. reasons for the said feeling) to sustain and advance the discourse (Turn 2).
2 points	• Interacts by responding intelligibly and coherently to interlocutor(s) for at least 50% of the time but exhibits some noticeable lapses in the expression of ideas. • Responds by revealing the nature of good news (Turn 1) and expressing how s/he feels about moving away (Turn 2) in a short, discrete utterance consisting of erroneous structures or mispronunciations impeding intelligibility/comprehensibility.
1 point	• Interacts by responding intelligibly and coherently to interlocutor(s) for at least 30% of the time. • Responds to the first question, with an attempt to share the nature of good news (Turn 1), followed by silence or lack of intelligible or coherent responses at the subsequent turns.
0 points	• Makes no attempt to respond or the response is unintelligible or unrelated to the topic. • Makes no or little attempt to sustain or advance the discourse at the first turn and thereafter, falling silent and/or providing incomplete responses.

Appendix C. Learner Perception Questionnaire Items used in the Current Study

System	Facilitating conditions (FC)	1. The instructions for completing the test were presented clearly. 2. The audio volume could be adjusted and was easy to hear. 3. I was able to take the test without being disturbed by surrounding noise. 4. The screen size was appropriate for test-taking. 5. The font size was appropriate for test-taking. 6. The resolution of illustrations and photographs was appropriate for the test. 7. I was able to access the automated scoring (results viewing system) without difficulty. 8. The time it took for the test score and feedback to be returned after the test was appropriate.
Task	Task appropriateness (TA)	1. The level of difficulty of the test was appropriate. 2. I was able to understand what content I was expected to respond with.
	Test-taking anxiety (TTA)	1. I feel anxious about taking automated-scoring speaking tasks such as these. 2. I worry that I might make an operational mistake that cannot be corrected. 3. I feel uneasy that a technical malfunction might occur during the task.
	Self-efficacy (SE)	1. I think I could complete automated-scoring speaking tasks such as these on my own if a manual were provided. 2. I think I could complete automated-scoring speaking tasks such as these on my own if no one else were present.
Feedback	Cognitive feedback (CF)	1. I think the test score and feedback accurately reflect my ability. 2. The feedback provided with the test score was clear and understandable. 3. Any additional information you would like to be included on the feedback page. (open-ended)

(Continued)

(Continued)

Feedback	Behavioral intention to use (BIU)	1. I would recommend automated-scoring speaking tasks such as these to others.
		2. I think I will use automated-scoring speaking tasks such as these frequently in the future.
		3. I intend to continue using automated-scoring speaking tasks such as these in the future.
		4. I think I can make effective use of the feedback returned with the test score after the test.
	Perceived usefulness (PU)	1. I think the feedback returned with the test score after the test is useful for guiding my future learning.
		2. I think the feedback returned with the test score after the test will contribute to improving my speaking ability.
		3. I think automated-scoring speaking tasks such as these can also be useful for preparing for other speaking tests.
		4. Reasons for finding the test score and feedback useful. (open-ended)

Appendix D. The Full List of Participants’ Responses to PU4 ‘Reasons for finding the Feedback Useful’

Theme	Participant comments
Diagnostic and specific feedback (n = 22)	The feedback clearly showed what I did well and what I still needed to improve. The weak points were described in detail. The reasons for receiving full marks (3 out of 3) were explained in detail. It clearly stated how my responses led to the score I received. I could see which parts I had failed to answer. Clear comments were written alongside my results. I was able to learn what was wrong with my pronunciation. The feedback described specifically what I could and could not do. The analysis of my responses was accurate. Both my strengths and weaknesses were written clearly. The questions I found difficult were pointed out. I realised which words and questions I was weak at. The results showed concretely what kind of answer was expected. The feedback showed where I needed to improve and what I did well. The specific feedback on what I had failed to do was helpful. My performance was described in detail. The reasons for each score were explained. The results included comments on both weak areas and good points. I found out exactly which parts I was lacking in. My strong and weak areas were clearly indicated. The feedback accurately identified the parts I lacked confidence in. The feedback matched my own awareness of my weaknesses.
Understanding one’s current level of English proficiency (n = 18)	I was able to understand my current speaking, listening, and reading comprehension skills. I understood my current level of English ability and received immediate results. (n = 2) I could know my real ability right away. I realised my speaking ability was higher than I had expected. I was able to confirm my English ability. I could view an analysis of my English performance. I was able to visualise my ability and have it evaluated objectively. I was able to find out how well I can speak English. I could see my actual speaking ability expressed numerically. I could confirm how I responded and what I had said.

(Continued)

(Continued)

Theme	Participant comments
Understanding one’s current level of English proficiency (n = 18)	I learned my current situation and level. I could know my ability more accurately than before. I realised that I still cannot speak English naturally. I could see my ability compared with clear scoring criteria. The test allowed me to view my English performance objectively. I found out which parts of my speaking were good or bad. I could understand my speaking ability in concrete terms.
Reflection and awareness for future learning (n = 14)	I was able to reflect on what specific items I was able to respond to. Looking at the scoring criteria influenced how I plan to study and my overall learning attitude. I thought I could apply the results to future tests. The feedback helped me understand what to work on next. I realised I should work harder on my speaking. It motivated me to think about how to improve my study habits. The comments encouraged me to reflect on my learning attitude. I learned that when talking with foreigners, it is important not to remain silent. The feedback helped me reconsider how I approach English learning. I thought I could make use of it for future improvement. It influenced how I will prepare for speaking tasks from now on. I became aware of what I need to study more. I felt that I need to practise responding more quickly. It reminded me of how I performed and what I need to change next time.
Clarity, transparency, and fairness of scoring (n = 11)	I was satisfied with the explanation of the scoring. The reasons for my speaking test score were clear. The results were easy to understand and the scoring criteria were presented. The reasons for the results were clearly indicated. The scoring and explanations were fair and convincing. The analysis and scoring were accurate and objective. The feedback clearly explained the rationale behind each score. It was easy to understand how my answers were evaluated. I trusted the feedback since it matched my self-evaluation. I could understand the basis for the evaluation right away. The scoring appeared systematic and transparent.
Practical usefulness of the feedback (n = 9)	I can definitely make use of it in the future. The feedback provided concrete advice for improving my English. The results will be useful for future English study and tests.

(Continued)

(Continued)

Theme	Participant comments
Practical usefulness of the feedback (n = 9)	I learned specific expressions I can apply next time.
	The results gave me actionable guidance for learning.
	The comments were specific enough to use as reference material.
	I can apply the feedback to future language learning.
	The test helped me learn how to study more effectively.
Motivation and encouragement (n = 7)	The feedback provided practical advice that can be used in daily learning.
	It made me want to work harder on English.
	It helped increase my motivation.
	I felt I should speak more confidently and clearly.
	The feedback encouraged me to continue improving.
	It gave me confidence in my speaking ability.
	It motivated me to make efforts to improve.
	I realised that improvement is possible through practice.