

Assessing Speaking on the Duolingo English Test



Duolingo Research Report DRR-23-03
Original Version: June 12, 2023 (15 pages)
Current Version: June 20, 2025 (10 pages)
<https://englishtest.duolingo.com/research>

Yena Park*, Ramsey Cardwell*, Sarah Goodwin*, Ben Naismith*, Geoffrey T. LaFlair*, Kevin P. Yancey*, Xiaowan Zhang*

Abstract

This paper presents the Duolingo English Test’s speaking construct, situated within the Duolingo English Test assessment ecosystem (Burstein et al., 2022). We describe how the Duolingo English Test defines, operationalizes, and measures speaking through various speaking-related item types. The operationalization and measurement of the speaking construct includes the item-type design process and automated item generation processes.

Last Update: June 2025

Keywords

Duolingo English Test, speaking assessment, testing speaking

Contents

1	Introduction	2
2	Speaking in the Duolingo English Test Ecosystem	2
2.1	The Language Assessment Design Framework	2
2.1.1	Construct definition	2
2.1.2	Task Types	3
2.1.3	Automated Item Generation	5
2.1.4	Test-Taker Readiness Materials and Practice Tests	6
2.2	The Expanded Evidence-Centered Design Framework	6
2.2.1	The Task Model	7
2.2.2	The Observational-Evidence Model	7
2.2.3	The KSA (Knowledge, Skills, and Abilities) Model	8
2.3	The Computational Psychometric Framework	8
2.4	The Test Security Framework	8
3	Conclusion	8
4	References	9

*Duolingo, Inc.
Email: englishtest-research@duolingo.com

1 Introduction

The Duolingo English Test is a measure of English language proficiency for communication and use in English-medium settings (Cardwell et al., 2023). It assesses the ability to use the independent language skills of speaking, writing, listening, and reading as well as the integrated skills of literacy (reading and writing), comprehension (listening and reading), production (speaking and writing), and conversation (listening and speaking).

The capacity to speak a language is an integral aspect of overall proficiency (Luoma, 2004). This skill is essential for both social interaction and academic engagement. In educational settings, the ability to speak a language facilitates participation in discussions, presentations, and interpersonal communication with professors and peers (Rosenfeld et al., 2001)*.

To assess the constructs of speaking and interaction, The Duolingo English Test includes five different tasks:

- Interactive Speaking
- Speak about the Photo
- Read, Then Speak
- Speaking Sample
- Interactive Listening

All tasks here contribute to the Conversation subscore, with the first four tasks contributing additionally to the speaking subscore and the Production subscore. The Speaking Sample task, prior to May 2023, was an unscored task, meaning that responses to the task were not evaluated and did not contribute to any of the subscores or the overall score. As of May 1, 2023, the Speaking Sample task is scored and contributes to the overall score as well as the speaking subscores. The rationale for beginning to score the Speaking Sample was to (1) improve the construct coverage of speaking and (2) increase the reliability with which speaking ability is measured, all while maintaining the test length; this change mirrors the decision to score the Writing Sample task (see Goodwin et al., 2025).

The speaking tasks of the Duolingo English Test are unique in that they are digital-first, meaning that they are created, administered, and scored in a way that takes advantage of state-of-the-art technology. The items are created automatically using automatic item generation techniques; the responses are transcribed using automatic speech recognition and scored using automated scoring systems. Additionally, test takers' responses to the Speaking Sample task are made available to the institutions that test takers apply to, to aid holistic decision-making by providing a demonstration of test takers' speaking proficiency.

This paper situates the speaking construct of the Duolingo English Test in the assessment ecosystem (Burstein et al., 2022), spanning all participating frameworks from the Language Assessment Design Framework to the Test Security Framework.

2 Speaking in the Duolingo English Test Ecosystem

The Duolingo English Test ecosystem is a combined network of theoretical frameworks in language assessment design, measurement, and test security processes (Burstein et al., 2022). The ecosystem guides assessment design and the collection of evidence for the test's digitally-informed chain of inferences. This paper will discuss how speaking is represented on the Duolingo English Test using the assessment ecosystem frameworks.

2.1 The Language Assessment Design Framework

The Language Assessment Design Framework encompasses the design of tasks that elicit construct-relevant behaviors. It consists of five key components: construct definition, test item design, item generation and scoring, evidence-specification, and test-taker readiness materials.

2.1.1 Construct definition

A theoretical construct definition details the facets/components of an ability to be assessed. An explicit construct definition guides test development by ensuring test content and scoring sufficiently cover the target construct given the intended score use. There are multiple approaches to defining the construct of speaking. Speaking can be modeled through its underlying process of speech production: conceptualization, formulation, and articulation Kormos (2006). It can also be defined through the list of microskills and macroskills that support oral production (Brown, 2004, p. 142). Microskills span the range of linguistic repertoires from acoustics to syntax, and macroskills include strategies and nonverbal cues.

*We acknowledge that the language abilities encompass signed languages but they are not discussed here as they are not relevant to speaking on the Duolingo English Test

Another perspective conceptualizes speaking through the interaction of language competence and strategic competence in the context of target-language use (TLU) tasks (Bachman & Palmer, 2010). Language competence in Bachman and Palmer (2010) comprises both organizational and pragmatic knowledge. Organizational knowledge encompasses control of formal elements of language at the phonological, syntactical, lexical, and discourse levels. Pragmatic knowledge works in tandem with organizational knowledge to imbue the linguistic output with meanings and intentions and/or to contextualize the linguistic output to a particular use situation.

In addition to describing the language abilities and processes involved in speaking, the construct definition can also include specific TLU tasks that activate and require these abilities and processes, thereby reflecting the contextual nature of speaking. A needs analysis has been the typical means through which TLU tasks have been identified (Bachman & Palmer, 1996, 2010; Norris et al., 1998). Several comprehensive needs analyses for English-medium university settings have been published (Ferris & Tagg, 1996a, 1996b; Rosenfeld et al., 2001), but these do not necessarily reflect the recent surge in technology-mediated communication in university settings (Burstein et al., 2022). Notwithstanding the need to understand the effect of technology-mediated settings, speaking tasks that have been shown to be important in English-medium university settings include but are not limited to: asking questions, making comments and suggestions, participating in class discussions, presenting in class; summarizing information; giving directions or instructions; describing objects; giving and supporting opinions; making comparisons/contrasts; developing or structuring hypotheses; explaining or informing (Rosenfeld et al., 2001).

The DET's construct definition is informed primarily by Bachman and Palmer (2010) and The Council of Europe (2020). The Common European Framework of Reference for Language (CEFR) (The Council of Europe, 2020) lists five oral production activities of potential relevance to the construct definition of speaking on the DET. Three activities (describing experience, giving information, and putting a case) fall under the larger umbrella of sustained monologue, while two activities (public announcements and addressing audiences) are more specialized genres of speaking that include the ability to consider the audience and to handle questions. Additionally, the CEFR manual has a separate list of oral interaction activities that can also be a part of the DET's target speaking construct. The oral interaction activities include understanding an interlocutor, conversation, informal and formal discussion, goal-oriented cooperation, obtaining goods and services, information exchange, interviews, and telecommunications. More information about how the Duolingo English Test assesses interaction with Interactive Listening and Interactive Speaking can be found in LaFlair et al. (2023) and Attali et al. (2025).

2.1.2 Task Types

The speaking tasks on the Duolingo English Test map largely to the oral production activities outlined by CEFR and elicit discourse across all CEFR levels of proficiency from A1 through C2.

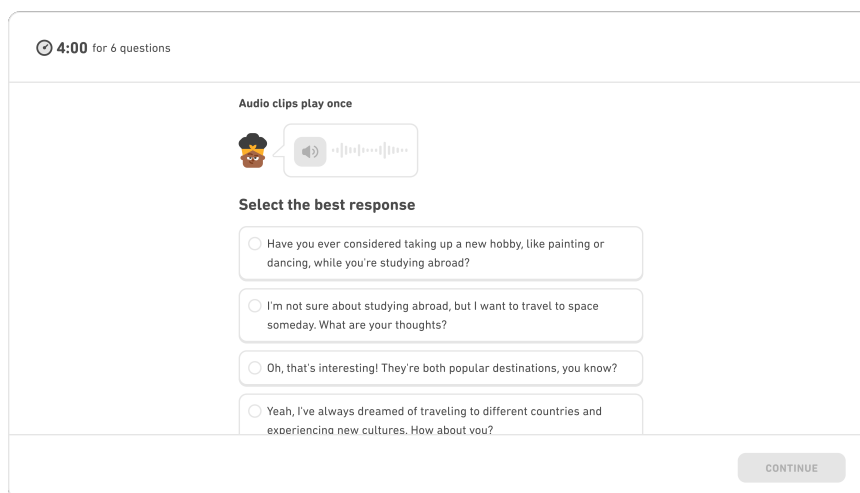


Figure 1. A screenshot of the Interactive Listening task on the Duolingo English Test

The Interactive Listening and Interactive Speaking are both new tasks introduced to the test with a focus on interaction. The Interactive Listening task (see Figure 1) requires test takers to participate in a situationally driven conversation in a university setting (LaFlair et al., 2023). By asking test takers to select the best response to an aurally delivered turn, the task indirectly assesses test takers' spoken abilities to sustain informal and formal discussion, obtain goods and services, exchange information, and participate in an interview.

The Interactive Speaking task (see Figure 2, Figure 3, and Figure 4) provides a direct assessment of oral interaction skills (Attali et al., 2025). It engages test-takers in a spontaneous conversation with an avatar interlocutor that asks 6-8 questions about topics ranging from personal, familiar matters to complex and abstract subjects. The test-taker has up to 35 seconds to answer each question. After they finish answering a question, their

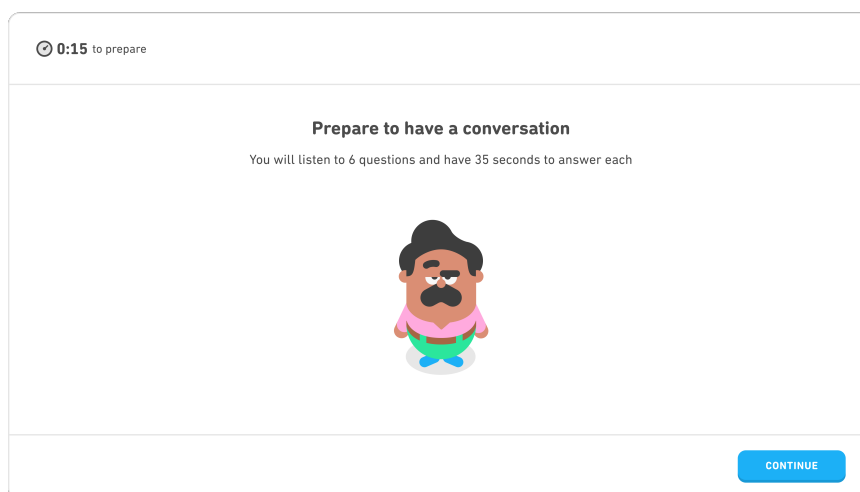


Figure 2. A screenshot of the Interactive Speaking task on the Duolingo English Test - Preparation

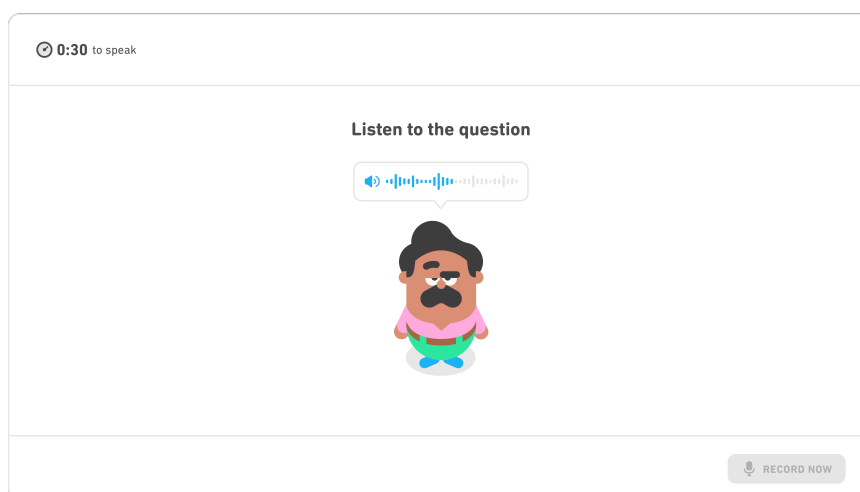


Figure 3. A screenshot of the Interactive Speaking task on the Duolingo English Test - Listening

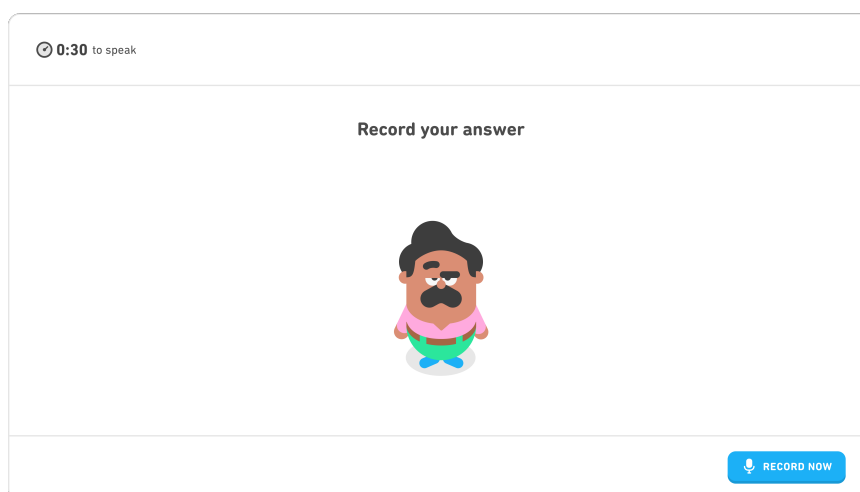


Figure 4. A screenshot of the Interactive Speaking task on the Duolingo English Test - Speaking

response is transcribed and scored in real time to inform the selection of follow-up questions. Designed to simulate dialogic communication, this task assesses interactional abilities such as understanding an interlocutor and participating in conversations or discussions.

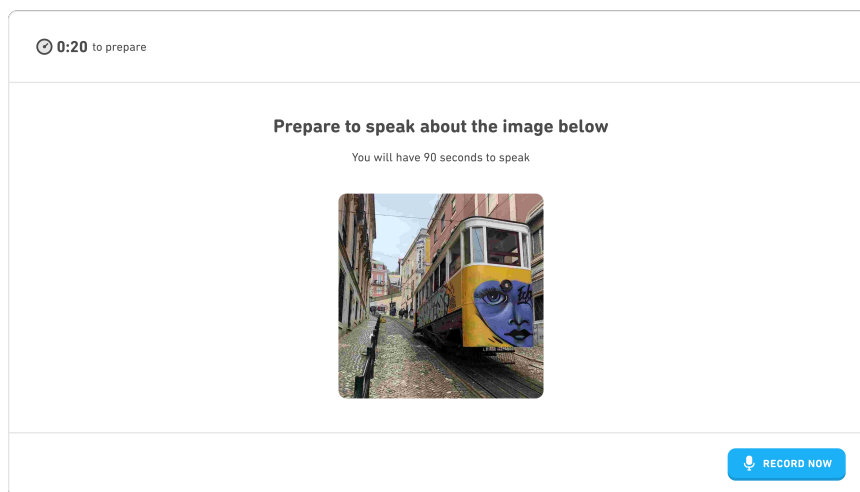


Figure 5. A screenshot of the Speak about the Photo task on the Duolingo English Test

Speak about the Photo (see Figure 5) requires test takers to describe the photo in 90 seconds, after 20 seconds of planning. Test takers engage in CEFR-aligned oral production activities, such as giving information by describing objects, to produce responses.

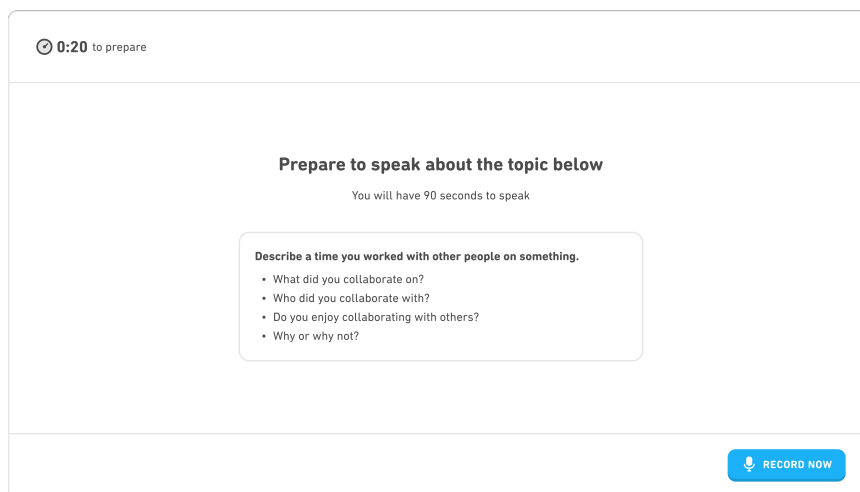


Figure 6. A Screenshot of the Read, Then Speak task on the Duolingo English Test

Read, Then Speak (see Figure 6) requires test takers to read a prompt, prepare for 20 seconds, and give a response in 90 seconds. The task provides an overarching question statement, with bullet points expanding the question statement that provide scaffolding for test takers. The test takers are asked to provide descriptions and narratives, give information, and sustain an argument, which are part of the CEFR oral production activities.

The Speaking Sample (see Figure 7) requires test takers to read a prompt, prepare for 30 seconds, and then talk for 3 minutes. Given that test-taker responses to this task are shared to test users (namely, the admissions officers) in the form of a video recording, the task invokes test takers' ability to consider the audience when delivering a sustained monologue, which is part of the oral production activities outlined in the CEFR manual.

Duolingo English Test tasks are also designed to mitigate the influence of construct-irrelevant sociocognitive factors Weir (2005). Each free-response speaking task (Speak about the Photo; Read, Then Speak; and Speaking Sample), has 20 to 30 seconds of planning time to mitigate any adverse effects on the performance due to intrapersonal and neurological factors, such as confidence and working memory (Elder et al., 2002; Wigglesworth & Elder, 2010). Additional planning time is provided as an accommodation to test takers with relevant disabilities (Cardwell et al., 2023).

2.1.3 Automated Item Generation

The prompts for speaking tasks are generated automatically using large-language models and go through automated filters and human reviews.

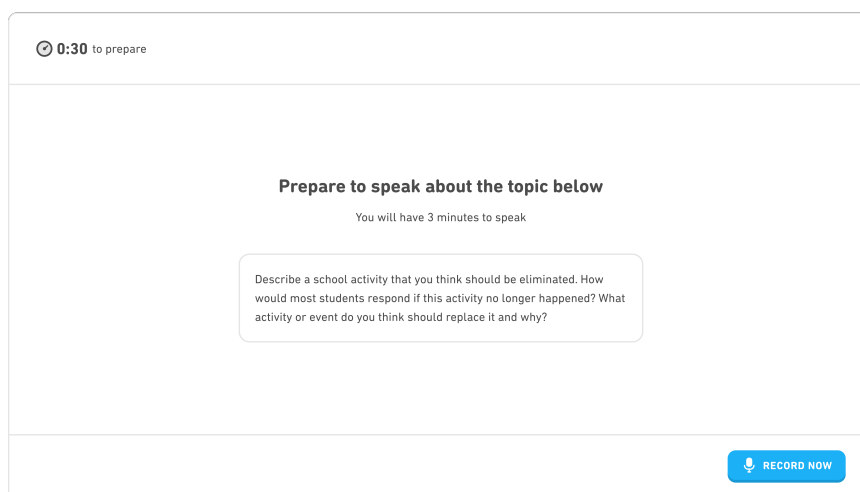


Figure 7. A Screenshot of the Speaking Sample task on the Duolingo English Test

Prompts are generated to exemplify two rhetorical functions: narrative and persuasive. The latter encompasses a variety of genres including expository, argumentative, comparative, and evaluative functions (Butt et al., 2003). The prompts are generated based on a number of requirements, including length and lexical frequency, and are filtered automatically based on multiple criteria related to topic (e.g., if a topic is engaging, broad, thought-provoking, and accessible to people from all backgrounds) and language (e.g., if the prompt is sufficiently clear and coherent).

The prompts then go through two stages of human review. Each item is reviewed by at least four people—first, in the item quality review stage where AI-generated prompts are reviewed and edited by item writers to ensure adherence to internal standards for item development; and second, in the fairness and bias (reduction) stage where prompts are reviewed by three reviewers separately for potential bias in the content that may interfere with the fairness of the test. Any item that is marked for revision by a fairness and bias reviewer is returned to an item quality reviewer for further edits. Finally, items that survive the reviews go through large-scale piloting to ensure sound psychometric qualities.

The item generation and human review processes for interactive tasks (i.e., Interactive Speaking and Interactive Listening) are slightly more complex than those for monologic speaking tasks. Please see LaFlair et al. (2023) for details about Interactive Listening and Attali et al. (2025) for more about Interactive Speaking.

2.1.4 Test-Taker Readiness Materials and Practice Tests

In addition to the test readiness guide, [Duolingo English Test: Official Guide for Test Takers](#), the Duolingo English Test offers free unlimited practice tests that simulate the test-taking experience to provide opportunities for test takers to familiarize themselves with the test format and receive an estimate of their overall score on the full-length test. We are regularly updating our offerings of test readiness materials to provide more resources for test takers to feel confident when taking the DET.

2.2 The Expanded Evidence-Centered Design Framework

The Expanded Evidence-Centered Design (eECD) Framework provides a comprehensive and structured approach to assessment design and development with implications for learning Mislevy et al. (2003). It is composed of three models: the Task Model; the Observational-Evidence Model; and the Knowledge, Skills, and Abilities (KSA) Model. The Task Model outlines the task features that elicit the underlying construct; in the context of the speaking construct, this would include design features of the speaking tasks that may have implications for elicited behaviors related to speaking. The Observational-Evidence Model defines the relationship between the elicited performance and the latent competency; more specifically, the Observational-Evidence Model guides decisions on what features to extract from spoken responses and helps define their relationship to the construct of speaking and its subconstructs. The KSA Model specifies the target construct, namely English proficiency; the KSA Model delineates how the speaking scores from the Observational-Evidence Model relates to subscores and the overall score.

The focus in this section is limited to tasks that contribute to the speaking subscore.

2.2.1 The Task Model

The Task model guides decisions about how speaking items are developed and administered. This is important as each decision has the potential to introduce construct-irrelevant variance that may influence how test takers respond to prompts and subsequently the scores and the interpretations we make from the scores. Here we highlight some task features of speaking tasks on the Duolingo English Test.

All speaking tasks (except Interactive Speaking) are accompanied by dedicated pre-task planning time not only to mitigate construct-irrelevant sociocognitive factors, but also to help facilitate test takers' oral production (Robinson, 2001; Skehan, 2009) (for Interactive Speaking, planning time is built into the question response time). Extended response time (up to 3 minutes for Speaking Sample, and up to 90 seconds for other tasks) is provided to encourage opportunities for online planning and monitoring (Weir et al., 2003). Read, Then Speak provides scaffolding by means of bullet points that extend the main prompt for test takers to optionally utilize (Weir et al., 2003).

Any speaking item that has been administered to a test taker will not be presented to the same test-taker in their subsequent test administrations for 30 days. This practice is consistent across all item types on the Duolingo English Test. While acknowledging the benefit of task repetition for improvement in test taker performance (Bui et al., 2019), the rationale behind the limitation on item re-administration is to promote fairness for all test takers by eliciting their most recent, current oral performance independent of potential practice effects or pre-knowledge of items.

As described above, a variety of tasks is included on the test to assess the construct of speaking ability from different perspectives. These tasks elicit monologic and dialogic speech samples addressing a wide range of topics and communicative purposes. The use of multiple task types is intended to provide a more accurate and comprehensive picture of speaking proficiency as it relates to real-world language use in university settings (Chapelle et al., 2008).

2.2.2 The Observational-Evidence Model

The Observational-Evidence Model dictates what observable construct-relevant feature measures to extract from raw data and how they relate to the construct. The speaking grader on the Duolingo English Test specifies multiple features that represent the construct of speaking. Numerical values on these features are calculated and combined to form a grade for a speaking item. There is a separate grading model for each speaking item type. Speaking grades for four speaking items (Interactive Speaking; Speak about the Photo; Read, Then Speak; and Speaking Sample) are then combined to form a score for speaking.

Spoken responses on the Duolingo English Test are graded along the [task-specific, holistic scale rubrics](#) that are publicly available. These rubrics assess speaking subconstructs of content, discourse coherence, lexis, grammar, fluency, and pronunciation. The rubrics are on the CEFR scale from A1 through C2 (six levels), with level descriptors informed by the CEFR descriptors as well as other speaking assessment rubrics. Table X presents example dimensions measured by each subconstruct.

Subconstruct	Example Dimensions
Content	Task achievement, relevance, effect on the reader, appropriacy of style, development
Discourse coherence	Clarity, cohesion, structure, progression of ideas, appropriacy of format
Lexis	Lexical diversity, lexical sophistication, word choice, word formation, spelling, error severity
Grammar	Range of structures, grammatical complexity, error frequency, error severity, appropriacy
Fluency (speaking only)	Speed, chunking, breakdowns, repairs
Pronunciation (speaking only)	Intelligibility, individual sounds, word stress, sentence stress, intonation

Within each subconstruct, features of the relevant dimensions are extracted and computed for each response. For instance, the fluency subconstruct can further be categorized into dimensions measuring speed fluency, breakdown fluency, and repair fluency, constructing granular representation of utterance fluency (Housen & Kuiken, 2009; Skehan, 2003). Features related to the Lexis subconstruct measure the breadth and depth of lexical knowledge, word choice, error severity, and mechanics such as spelling (Bulté & Housen, 2012; Skehan & Foster, 2012). Grammar is represented by the dimensions of structure variation, grammatical complexity, and accuracy of spoken responses (e.g., error frequency, error severity, appropriacy). The pronunciation subconstruct measures both segmental and suprasegmental speech aspects including pronunciation of individual sounds, word/sentence stress, intonation, and intelligibility of oral responses. Within the subconstruct of content, features related to task achievement, topic relevance, reader orientation, style appropriateness, and topic development are computed for each response. Some examples are differential word use features (Attali, 2011) and relevance features for content and cohesion features for discourse management. Discourse coherence is operationalized through features related to clarity, cohesion, structure, idea progression, and format appropriacy.

The speaking scores based on the features outlined above are supported by concurrent validity evidence. There are strong correlations of DET speaking scores to speaking subscores of other high-stakes English proficiency tests, supporting the claim that the construct underlying DET

speaking scores and other tests' speaking subscores is similar (for specific numbers, see [Cardwell et al., 2023](#)). This claim can be extended to argue that DET speaking scores are indicative of speaking proficiency. Moreover, the speaking scores, all derived and computed automatically by the automatic scoring engine, correlate to human evaluations at a high level (e.g., [Isbell et al., 2024](#)). These high correlations provide support for the use of the automated scoring engine of the DET.

2.2.3 The KSA (Knowledge, Skills, and Abilities) Model

The KSA model governs how speaking scores are related to oral proficiency. The speaking features specified in Section 2.2.2 have been shown to differentiate test takers by proficiency level and are used frequently in second language acquisition research as components of speaking proficiency ([Fan & Yan, 2020](#); [Kang & Yan, 2018](#); [Sato, 2012](#); [Tavakoli & Skehan, 2005](#)).

2.3 The Computational Psychometric Framework

The Computational Psychometric Framework guides decisions related to statistical and machine learning modeling. Brief descriptions regarding transcribing, feature weighting, and scaling are provided.

All test-taker speech is transcribed using Whisper, a robust speech recognition system that approaches human-level accuracy ([Radford et al., 2021](#)). Transcription accuracy is evaluated for gender, device, and first language bias by comparing group differences on indices for transcription accuracy such as word error rates, match error rates (probability of a given match being incorrect), and word information lost (an approximation of relative information lost based on mutual information).

Transcription-based and acoustic features are extracted from oral responses using computational methods. These features are later fed into a regression model to grade responses on a norm-referenced scale with grades strongly correlated with the ratings assigned by human examiners against the holistic rubrics mentioned earlier. Grades of all four speaking tasks are aggregated into an overall speaking subscore. The speaking scores are then scaled to produce numerical values that represent how much speaking contributes to the subscores and the overall scores. More information on subscores and the overall score can be found in [Cardwell et al. \(2023\)](#) and [LaFlair \(2020\)](#). The addition of Speaking Sample to the scored test content led to a slight increase of the test-retest reliability of speaking scores and the Conversation subscore.

2.4 The Test Security Framework

The Test Security Framework contributes to the validity argument by providing evidence that ensures the speech samples represent the original and unaided effort of the individual with whom the scores are associated. This is ensured through multiple security measures in place, including the proprietary Duolingo English Test desktop application, gaze detection, plagiarism detection, human proctoring, and the scrutinization of extreme discrepancies between performance on speaking items and non-speaking items ([Duolingo English Test, 2021](#); [LaFlair et al., 2022](#)).

3 Conclusion

The paper situates the Duolingo English Test's speaking construct within the assessment ecosystem and illustrates how speaking is conceptualized, operationalized, and scored. The Duolingo English Test assesses the construct of speaking directly through prompts that elicit speech pertaining to multiple different rhetorical purposes relevant to university settings and indirectly through selected responses in a simulated conversation task. Item types on the Duolingo English Test tap into multiple subconstructs of speaking, from language control to discourse management. Beginning to score the Speaking Sample task (which was unscored prior to May 2023) serves to strengthen the construct of speaking through increased reliability, longer responses, and varied audiences.

4 References

- Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B., & von Davier, A. A. (2019). The expanded evidence-centered design (e-ECD) for learning and assessment systems: A framework for incorporating learning goals and processes within assessment design. *Frontiers in Psychology, 10*. <https://doi.org/10.3389/fpsyg.2019.00853>
- Attali, Y. (2011). A differential word use measure for content analysis in automated essay scoring. *ETS Research Report Series, 2011*(2), i–19. <https://doi.org/10.1002/j.2333-8504.2011.tb02272.x>
- Attali, Y., Runge, A., Church, J., Park, Y., & Goodwin, S. (2025). *Generative-AI automatic item generation for large-scale language assessments*. Presented at the National Council on Measurement in Education Conference, Denver, CO. https://docs.google.com/presentation/d/1UwHOpAwMUDH6LBTfL0V2_QqLsggrYLaQAQnKtnBX3Y/edit?slide=id.g1e01014f019_0_292#slide=id.g1e01014f019_0_292
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.
- Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Longman.
- Bui, G., Ahmadian, M. J., & Hunter, A.-M. (2019). Spacing effects on repeated L2 task performance. *System, 81*, 1–13. <https://doi.org/10.1016/j.system.2018.12.006>
- Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA* (pp. 21–46). John Benjamins. <https://doi.org/10.1075/llt.32.02bul>
- Burstein, J., LaFlair, G. T., Kunnan, A. J., & von Davier, A. A. (2022). *A theoretical assessment ecosystem for a digital-first assessment—The Duolingo English Test* (Duolingo Research Report DRR-22-01; pp. 1–32). Duolingo. <https://go.duolingo.com/ecosystem>
- Butt, D., Fahey, R., Feez, S., Spinks, S., & Yallop, C. (2003). *Using functional grammar: An explorer's guide*.
- Cardwell, R., Naismith, B., LaFlair, G. T., & Nydick, S. (2023). *Duolingo English Test: Technical Manual* (pp. 1–40). Duolingo. <https://go.duolingo.com/dettechnicalmanual>
- Chapelle, C., Enright, M., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the test of english as a foreign language*. Routledge.
- Duolingo English Test. (2021). *Duolingo English Test: Security, proctoring, and accommodations* (pp. 1–9) [Duolingo Research Report]. Duolingo. <https://go.duolingo.com/securitywhitepaper>
- Elder, C. A., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing, 19*(4), 347–368. <https://doi.org/10.1191/0265532202lt235oa>
- Fan, J., & Yan, X. (2020). Assessing speaking proficiency: A narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in Psychology, 11*. <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00330>
- Ferris, D., & Tagg, T. (1996a). Academic listening/speaking tasks for ESL students: Problems, suggestions, and implications. *TESOL Quarterly, 30*(2), 297–320. <https://doi.org/10.2307/3588145>
- Ferris, D., & Tagg, T. (1996b). Academic oral communication needs of EAP learners: What subject-matter instructors actually require. *TESOL Quarterly, 30*(1), 31–58. <https://doi.org/10.2307/3587606>
- Goodwin, S., Attali, Y., LaFlair, G. T., Park, Y., Runge, A., Davier, A. von, & Yancey, K. P. (2025). *Duolingo English Test—Writing Construct* (Duolingo Research Report DRR-22-03; pp. 1–15). Duolingo. <https://go.duolingo.com/scored-writing>
- Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 461–473. <https://doi.org/10.1093/applin/amp048>
- Isbell, D. R., Crowther, D., & Nishizawa, H. (2024). Speaking performances, stakeholder perceptions, and test scores: Extrapolating from the duolingo english test to the university. *Language Testing, 41*(2), 233–262. <https://doi.org/10.1177/02655322231165984>
- Kang, O., & Yan, X. (2018). Linguistic features distinguishing examinees' speaking performances at different proficiency levels. *Journal of Language Testing & Assessment, 1*(1), 24–39. <https://doi.org/10.23977/langta.2018.11003>
- Kormos, J. (2006). *Speech production and second language acquisition*. Erlbaum. <https://doi.org/10.4324/9780203763964>
- LaFlair, G. T. (2020). *Duolingo English Test: Subscores* (Duolingo Research Report DRR-20-03; pp. 1–16). Duolingo. <https://go.duolingo.com/subscorewhitepaper>
- LaFlair, G. T., Langenfeld, T., Baig, B., Horie, A. K., Attali, Y., & von Davier, A. A. (2022). Digital-first assessments: A security framework. *Journal of Computer Assisted Learning, 38*(4), 1077–1086. <https://doi.org/10.1177/0265532211421162>
- LaFlair, G. T., Runge, A., Attali, Y., Park, Y., Church, J., & Goodwin, S. (2023). *Interactive Listening—The Duolingo English Test* (Duolingo Research Report DRR-23-01; pp. 1–17). Duolingo. <https://go.duolingo.com/interactive-listening-whitepaper>
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. The MIT Press.
- Luoma, S. (2004). *Assessing Speaking*. Cambridge University Press.
- Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), 1–29.
- Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments: Technical report*. Second Language Teaching & Curriculum Center, University of Hawai'i at Manoa.

- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2021). *Robust speech recognition via large-scale weak supervision* (pp. 1–28). OpenAI. <https://cdn.openai.com/papers/whisper.pdf>
- Robinson, P. (2001). Task-based language learning: A review of issues. *Language Learning*, 61(s1), 1–36. <https://doi.org/10.1111/j.1467-9922.2011.00641.x>
- Rosenfeld, M., Leung, S., & Oltman, P. K. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels* (RM-01-03; pp. 1–104). Educational Testing Service. <https://www.ets.org/Media/Research/pdf/RM-01-03.pdf>
- Sato, T. (2012). The contribution of test-takers’ speech content to scores on an English oral proficiency test. *Language Testing*, 29(2), 223–241. <https://doi.org/10.1177/0265532211421162>
- Skehan, P. (2003). Task-based instruction. *Language Teaching*, 36(1), 1–14. <https://doi.org/10.1017/S026144480200188X>
- Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510–532. <https://doi.org/10.1093/applin/amp047>
- Skehan, P., & Foster, P. (2012). Complexity, accuracy, fluency and lexis in task-based performance: A synthesis of the Ealing research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA* (pp. 199–220). John Benjamins. <https://doi.org/10.1075/llt.32.09fos>
- Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). John Benjamins. <https://www.jbe-platform.com/content/books/9789027294647-llt.11.15tav>
- The Council of Europe. (2020). *Common European Framework of Reference for languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing. <https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4>.
- Weir, C. (2005). *Language testing and validation: An evidence-based approach*. Palgrave MacMillan.
- Weir, C., O’Sullivan, B., & Horai, T. (2003). *Exploring difficulty in speaking tasks: An intra-task perspective* (No. 5; pp. 1–42). IELTS. https://www.ielts.org/-/media/research-reports/ielts_rr_volume06_report5.ashx
- Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly*, 7(1), 1–24. <https://doi.org/10.1080/15434300903031779>