Assessing Speaking on the Duolingo English Test


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.

Keywords
Duolingo English Test, speaking assessment, testing speaking

* Duolingo, Inc.

Corresponding author:
Yena Park
Duolingo, Inc. 5900 Penn Ave
Pittsburgh, PA 15206, USA
Email: englishtest-research@duolingo.com

© 2023 Duolingo, Inc. All rights Reserved.
Contents

1 Introduction 3

2 Speaking in the Duolingo English Test Ecosystem 4
  2.1 The Language Assessment Design Framework 4
  2.1.1 Construct Definition 4
  2.1.2 Task Types 5
  2.1.3 Automated Item Generation 9
  2.1.4 Test-Taker Readiness Materials and Practice Tests 9
  2.2 The Expanded Evidence-Centered Design Framework 9
  2.2.1 The Task Model 10
  2.2.2 The Observational-Evidence Model 10
  2.2.3 The KSA (Knowledge, Skills, and Abilities) Model 11
  2.3 The Computational Psychometric Framework 11
  2.4 The Test Security Framework 12

3 Conclusion 12

4 References 12
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 language skills that are required for literacy, comprehension, production, and conversation.

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 construct of speaking, The Duolingo English Test includes six different tasks:

1. Read Aloud
2. Interactive Listening
3. Speak about the Photo
4. Read, Then Speak
5. Listen, Then Speak
6. Speaking Sample

All speaking tasks listed here contribute to the Conversation subscore; additionally, the latter four speaking tasks are free-response tasks that contribute to 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 Conversation and Production 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., 2022).

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.

*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
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.

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 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: 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 can be found in LaFlair et al. (2023).

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

The Read Aloud task requires test takers to read out loud a sentence presented in text in 60 seconds (see Figure 1). The task assesses the organizational knowledge–namely segmental/phonemic and suprasegmental properties that affect the intelligibility and fluency of speech production.

The Interactive Listening task is a new task that 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.
Speak about the topic below for 90 seconds.

Talk about studying science in school.
- What kind of science do students study?
- Is it important to study science?
- Why or why not?

Figure 4. A Screenshot of the Read, Then Speak task on the Duolingo English Test

Speak about the Photo requires test takers to describe the photo in 90 seconds, after 20 seconds of planning (see Figure 3). Test takers engage in CEFR-aligned oral production activities, such as giving information by describing objects, to produce responses.

Read, Then Speak 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.

Listen, Then Speak gives test takers 20 seconds to listen to an aural prompt and prepare a response. Test takers are allowed to listen to the question up to two times. Test takers are then given 90 seconds to speak their response to the prompt. The prompts require test takers to provide descriptions and narratives, give information, and/or sustain an argument. The aural format of prompt delivery allows for simulating conversations and provides opportunities for test takers to exhibit their ability to handle questions.

The Speaking Sample 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, Listen, 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,
Assessing Speaking on the DET

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.

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.

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 tasks 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 eliciting free-response production.
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 four task features of speaking tasks on the Duolingo English Test.

All speaking tasks are accompanied by 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). 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 be excluded from item selection in any 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.

Among four non-adaptive task types that assess speaking on the Duolingo English Test, only Listen, then Speak is administered twice; all the other tasks are administered only once. This task configuration maximizes the number of test tasks that simulate the context of conversations and “bias[es] for the best”—providing the grounds for test takers’ maximum success (Brown, 2004; Fox, 2004; Swain, 1984). Despite mixed empirical findings (see Kim & Tracy-Ventura, 2013), it has been argued that procedural repetition—repeating the same task but with different content—may facilitate some aspects of language proficiency while speaking (Bygate, 2001; Lynch & Maclean, 2000).

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 scoring model for each speaking item type. Speaking grades for five speaking items are then combined to form a score for speaking.

Spoken responses on the Duolingo English Test are graded along the dimensions of language (Fan & Yan, 2020) and content (Kang & Yan, 2018; Sato, 2012). Within the dimension of language, features related to fluency, vocabulary, grammar, and pronunciation are computed for each response. Fluency features can further be categorized into features measuring speed fluency, breakdown fluency, and repair fluency, constructing granular representation of utterance fluency (Housen & Kuiken, 2009; Skehan, 2003). Vocabulary features measure the breadth and depth of lexical knowledge, including the degree to which it is used correctly and appropriately in context (Bulté & Housen, 2012; Skehan & Foster, 2012). Grammar features measure the grammatical complexity and accuracy of spoken responses. Pronunciation features
measure intelligibility of oral responses. Within the dimension of content, features related to
topic relevance and discourse management 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.

The speaking scores based on the features outlined above are supported by convergent validity
evidence. There are moderate-to-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. 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 extracted from oral responses using computational
methods are then weighted. In other words, each feature is given a weight that denotes the
degree to which the feature contributes to the grade, signifying its importance in the construct of
speaking. The weights were determined based on a proxy for English ability—in other words, the
performance on the DET excluding the speaking portions—and CEFR-aligned rating data based
on human expert judgment.

The scaling parameters for speaking scores were derived in a similar manner. Similar weighting
procedures were applied to the grade for each item to derive an overall speaking score that
comprises five grades to five items. The speaking scores were 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


Assessing Speaking on the DET

accuracy, and fluency in SLA (pp. 21–46). John Benjamins. https://doi.org/10.1075/lllt.32.02bul


