

Assessing Listening on the Duolingo English Test



Duolingo Research Report DRR-23-02
Original Version: June 13, 2023 (16 pages)
Current Version: June 25, 2025 (10 pages)
<https://englishtest.duolingo.com/research>

Sarah Goodwin* and Ben Naismith*

Abstract

In this paper we describe how the language skill of listening is operationalized and measured on the Duolingo English Test (DET). This work is situated in the DET’s theoretical assessment ecosystem (Burstein et al., 2022), a set of evidence-based frameworks that reflect the iterative processes for assessment design, computational psychometrics, and test security. In this ecosystem, the Language Assessment Design Framework stipulates that the domain for tested constructs be described. To achieve this goal, the present paper is one in an ongoing series of skills construct whitepapers that describes the underpinnings for each language skill construct, in this case for listening (see also Park et al., 2022 for reading; Goodwin et al., 2025 for writing; LaFlair et al., 2025 for interactional competence; Park et al., 2025 for speaking). The paper first gives background information on the DET. We then describe the DET’s conceptualization of the second language listening construct using the multi-layered framework of Aryadoust and Luo (2022). Within this framework, we consider how subskills, cognitive processes, attributes (i.e., task and test-taker traits) contribute to the overall listening construct. We also exemplify how these different elements of listening are measured through the DET task types.

Last Update: June 2025

Keywords

Duolingo English Test, testing listening, interactive listening, integrated listening, listening assessment

Contents

1	Introduction	2
2	The Duolingo English Test	2
2.1	Dictation item type	2
2.2	Interactive listening item type	3
3	Theoretical background	4
3.1	General theoretical background	4
3.2	Listening theoretical background	4
4	Listening approaches and the DET	5
4.1	DET subskills	5
4.2	DET processes	6
4.3	DET attributes	6
5	Summary and future directions	7
6	References	9

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

1 Introduction

In this paper we describe how the language skill of listening is operationalized and measured on the Duolingo English Test (DET). This work is situated in the DET's theoretical assessment ecosystem (Burststein et al., 2022), a set of evidence-based frameworks that reflect the iterative processes for assessment design, computational psychometrics, and test security. In this ecosystem, the Language Assessment Design Framework stipulates that the domain for tested constructs be described. To achieve this goal, the present paper is one in an ongoing series of skills construct whitepapers that describes the underpinnings for each language skill construct, in this case for listening (see also Park et al., 2022 for reading; Goodwin et al., 2025 for writing; LaFlair et al., 2025 for interactional competence; Park et al., 2025 for speaking). The paper first gives background information on the DET. We then describe the DET's conceptualization of the second language (L2) listening construct using the multi-layered framework of Aryadoust and Luo (2023). Within this framework, we consider how subskills, cognitive processes, and attributes (i.e., task and test-taker traits) contribute to the overall listening construct. We also exemplify how these different elements of listening are measured through the DET task types.

2 The Duolingo English Test

The DET is a high-stakes, English language proficiency test with the primary use case being postsecondary admissions. The DET is a digital-first test in that, from its inception, it was developed as an online test to leverage all of the affordances offered by this medium. It is available from anywhere in the world at any time, and it can be taken in approximately an hour. As such, the DET prioritizes accessibility while still ensuring that test scores are valid, reliable, and fair. This combination of test characteristics is possible because the DET is a computer-adaptive test (CAT) which continually adjusts the difficulty of test items to more efficiently estimate test-taker ability. In terms of scoring, the DET reports an overall score and eight subscores, four for the independent language skills of *Speaking*, *Writing*, *Reading*, and *Listening*, and four for the integrated skills subscores (LaFlair, 2020) of *Literacy* (reading and writing), *Comprehension* (reading and listening), *Conversation* (speaking and listening), and *Production* (speaking and writing). Task types which assess listening therefore contribute to the Listening, Comprehension, and Conversation subscores. (See Naismith et al., 2025, for a detailed description of the DET.)

We recognize that the skill of listening, both in test and non-test situations, is inherently integrated with the other skills of speaking, writing, and reading. An integrated-skills approach contrasts with the conceptualization of language users' speaking, writing, reading, and listening as independent skills. Traditional assessment tasks used discrete-point tasks meant to break language knowledge down into component parts (e.g., Lado, 1961, 1964), but later pushes toward communicative language teaching and assessment (e.g., Canale & Swain, 1980) wove skill use together (Bachman, 1990; Cumming, 2014; see McNamara, 2000 for a comparison between discrete and integrative language tests). In the real world, language skills are not used in isolation but instead together; for example, when engaging in a conversation, one needs to listen, speak in response, and occasionally write to summarize the interaction (e.g., in a follow-up email). Thus, when operationalizing the measurement of listening proficiency, listening ability as part of integrated modalities such as speaking-listening ability should be included in construct definitions (Ockey & Wagner, 2018). (See Park et al., 2025 for more detail.) The use of integrated-skills tasks on language assessments means that items can measure people's ability to use language in meaningful ways as they would in natural communication. Integration is also efficient in that language users employ their skills in an interconnected fashion and more dynamically than if the skills were targeted discretely, reinforcing learning across skills (see e.g., Oxford, 2002). Maintaining a focus on listening as a distinct skill is crucial for ensuring that test-takers receive diagnostic feedback specific to this ability and for guaranteeing the listening construct is adequately represented in the overall proficiency score.

There are various task types on the DET, all of which are scored. In the first phase of the test, task types which are part of the computer-adaptive administration include Yes/No Vocabulary, Vocabulary in Context, C-test, Read Aloud, and Dictation. These task types efficiently provide information about ability, primarily in terms of receptive skills (listening and reading) and linguistic resources (especially vocabulary and grammar). In the second phase of the test, communicative skills mastery is assessed through a series of more authentic task types comprising Interactive Reading, Interactive Listening, Picture Description (writing), Interactive Writing, Writing Sample, Picture Description (speaking), Extended Speaking (audio prompt), Extended speaking (text prompt), and Speaking Sample. The two task types that assess listening directly are Dictation and Interactive Listening. For further details, please see the DET Technical Manual (Naismith et al., 2025).

2.1 Dictation item type

For DET dictation items, test takers listen to a segment of recorded speech and are allotted 60 seconds to type what they hear into a text box. The dictation item stimuli can be played a maximum of three times; that is, after the first audio play, two replays are permitted. Test takers click a button with a speaker icon to play the utterance again. Each stimulus is a complete sentence, containing at minimum one independent clause. The stimulus sentences contain no proper nouns and range from 3 to 20 words, with more difficult items tending to contain more and longer words. DET dictation stimuli are limited to one sentence to discourage test takers from summarizing (as opposed to a verbatim dictation). No printed text

for the audio input is shown on screen other than instructions for the task. Responses are graded on a scale from 0 to 1 based on text similarity, with 1 meaning the response is identical to the expected response.

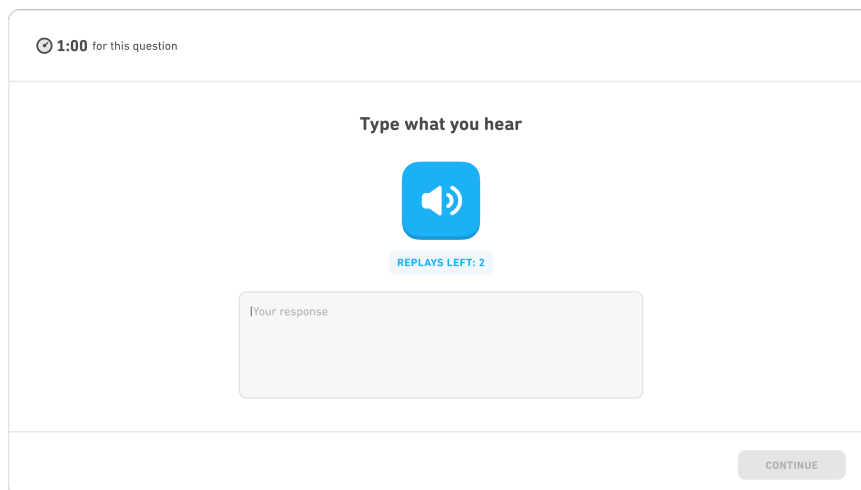


Figure 1. A screenshot of the Dictation task on the Duolingo English Test

2.2 Interactive listening item type

The Interactive Listening task has three interconnected parts: an audio scenario with comprehension questions, a conversation, and a writing task. In the interaction, the test taker plays the role of a university student, conversing with an avatar impersonating either another student or a professor. They first listen to a description of a scenario that sets the scene for the interaction, and they answer three related comprehension questions in the form of gap-fills. Next, they click a button with a play icon to play each utterance in the conversation; they can only listen to each utterance once. They select the most appropriate turn, given the context, by reading options in a printed-text online-chat format. Test takers then summarize the interaction in writing by typing into a text box. Each gap-fill scenario item is scored using a checklist-based rubric. The multiple-choice items use dichotomous scoring, and the written summary task is scored using a scoring model specific to that type of writing. More information about this item type can be found in LaFlair et al. (2025).

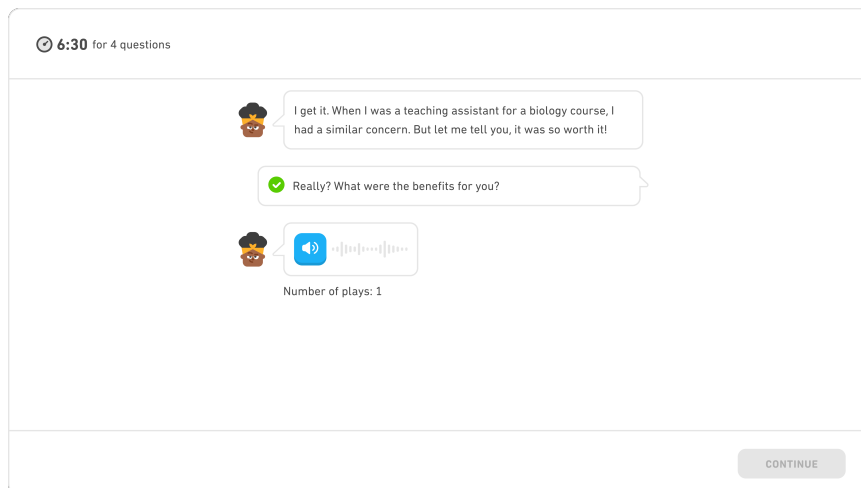


Figure 2. A screenshot of the Interactive Listening task on the Duolingo English Test

3 Theoretical background

In this section, we first describe the broad theoretical background which underpins the DET. We then focus on the theoretical background specific to the measurement of listening ability.

3.1 General theoretical background

The DET is grounded in the theoretical assessment ecosystem for digital-first assessments (Burstein et al., 2022). Woven through all of the theoretical ecosystem phases is a chain of inferences guiding the organization of validity evidence for score use. This chain of inferences includes domain description, item development, and extrapolation from results, with the test-taker experience considered throughout every part of the ecosystem. Test-taker experience includes characteristics such as low cost, accessibility of administration, and short administration time. The tests are designed from their incipient stages to prioritize test-taker access while also being pedagogically and psychometrically sound. Domain description, a main focus of the present whitepaper, highlights the subconstructs: the language skills the test assesses. Informing the language skill descriptions is the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001, 2020), an organized set of proficiency descriptors for use in language learning, teaching, and assessment. It conceptualizes language organized under broad categories of Reception, (comprehension of spoken and written language), and Production (speaking and writing ability).*

3.2 Listening theoretical background

The skill of listening has been operationalized in numerous ways. To describe the DET’s conceptualization of listening, we have opted to view it through the lens of the multicomponential framework outlined by Aryadoust and Luo (2023) because it takes as its basis a metaanalysis of 157 peer-reviewed listening papers in applied linguistics journals. In doing so, we aim to establish a broad theoretical basis for DET listening assessment which takes into consideration all of the major components of listening that are typically considered important in the assessment community. This goal aligns with the call to make listening assessments that are supported by empirical evidence and that have comprehensive construct coverage (Aryadoust & Luo, 2023; Taylor & Geranpayeh, 2011). In this framework (henceforth referred to as the A&L framework), three principal approaches to listening assessment are identified: Subskills, Processes, and Attributes as shown in Figure 3.



Figure 3. The three approaches to listening assessment (adapted from Aryadoust & Luo, 2023, p. 19)

In the subskills approach, the first-layer subskills in Figure 3 are consistent with previous research on linguistic knowledge and are essential for successfully achieving comprehension. (Second-layer subskills are not presented here.) These skills are often discrete and concrete in nature; that is, the ability to perform them can be externally observed and linked to tasks. For example, a test taker listening to an excerpt and answering comprehension questions about the main idea is intuitively a task that is an assessment of understanding global meaning. Unsurprisingly then,

*The CEFR also frames separate categories of Interaction (e.g., turn-taking and cooperating) and Mediation (e.g., adapting or simplifying text) which encompass features of both receptive and productive language use.

such skills are often specified within competency frameworks and targeted for assessment. It should be noted, however, that assessing knowledge of the sound system does not directly assess comprehension, but rather, enables other more global subskills (Aryadoust & Luo, 2023).

In contrast, the processes approach to listening assessment focuses on the cognitive processes that listeners employ to make sense of auditory input (Field, 2008). For example, top-down processing typically refers to the listener using background knowledge and context to understand the message being conveyed, whereas bottom-up processing often refers to decoding meaning by building up from smaller units (sounds → words → utterances, etc., Richards, 2008). Naturally, listeners use these and other processes simultaneously in a dynamic manner (Aryadoust, 2020), making it more challenging to tease apart or discretely assess compared to subskills.

Finally, attribute-based listening encompasses all aspects of listening assessment which affect performance, including features of the test or listeners. For example, test-related attributes include visual support for tasks, and listener-related attributes include demographic variables such as gender and age. As such, listening attributes are not the way in which a person listens, but rather, features of listening which must be taken into account in order to ensure fair and valid listening assessment.

4 Listening approaches and the DET

We now consider listening assessment on the DET through the lens of each of the three above approaches to describing listening.

4.1 DET subskills

The A&L framework does not specifically discuss the can-do descriptors of the CEFR. However, by their very nature, these descriptors are intended to illustrate discrete communicative abilities that language users can perform. We therefore include them here in our discussion of subskills, though attributes such as listening context and speaker roles are also a key element of CEFR descriptors. Specifically, we center our discussion on the connections between DET task types and listening skills as described in the CEFR. The CEFR references Personal, Public, Occupational, and Educational domains of language use, along with the conceptualization of pre-A1 to C2 levels, toward the goal of being able to describe a wide range of proficiency levels across most contexts. The framework focuses on teaching and learning additional languages (i.e., languages other than an individual's first or primary language), helping language professionals (e.g., teachers and researchers) to understand receptive subskills which are part of the listening construct. Although language assessment was not the intended primary purpose of the CEFR, this is now one of its main uses, and there have been numerous projects with efforts to align high-stakes language assessments to the CEFR (e.g., Alderson, 2002; Figueras & Noijons, 2009; Martyniuk, 2010).

In the CEFR, Reception and Interaction both pertain to listening ability. Here, Reception relates to descriptions of using reading or listening skills for tasks in relative isolation, in this case focusing on the listener (e.g., listening to announcements on public transportation). Reception is therefore the primary CEFR component assessed by the Dictation task. In contrast, interaction is conceptualized as more than just the reception or production of utterances, but involves using both to accomplish authentic communicative purposes, for example, taking turns, cooperating, and seeking clarification (Council of Europe, 2001, p. 14). Interaction is therefore a primary focus of the Interactive Listening task.*

The necessity for assessing different listening subskills is inherently tied to the various purposes for listening that language learners may encounter. For example, learners may need to comprehend broader or finer details in listening input. They also may need to understand a speaker's intention or rhetorical purpose—how or why someone said something. The DET listening tasks attempt to cover this spectrum of purposes; the Dictation task and the initial comprehension questions of the Interactive Listening task focus more on the details of listening input, whereas the conversation and summarization portions of the Interactive Listening task focus on elements such as speaker intention. Take for example the earlier illustrations of these two tasks (Figure 1 and Figure 2). For the Dictation task, the test taker listens to individual sentences that vary in difficulty depending on the linguistic features used in the utterance. In contrast, for the Interactive Listening task, the test taker is listening in order to understand the messages in the interaction, manage turns, and respond appropriately. We recognize that listening is done in the real world for more purposes than standardized testing situations always cover, but in this theoretical background, we focus on the CEFR listening purposes of gist, specific details, and rhetorical purpose.

One important listening subskill which is routinely practiced in language classrooms is *listening for gist*. Listening for gist refers to understanding of main ideas. It also involves comprehension of overall organizational structure of longer utterances matching with expectations (e.g., understanding that a lecture from a professor from start to finish may involve various predictable moves such as introducing the session, giving examples, or asking students if they have questions). This subskill is assessed as part of the Interactive Listening task. For example, in order to effectively summarize the conversation, it is not sufficient to have understood a few specific details. Rather, it is necessary to have understood the main ideas, the role of the two speakers, and the outcomes of the conversation.

*In addition, listening is assessed as a secondary construct as part of DET speaking-focused tasks; see Park et al. (2025) for further detail about these tasks.

In contrast, *listening for specific information* involves successfully pulling out a key, short piece of information from a longer utterance, like listening to a teacher to hear when an assignment is due, or listening to a public announcement for important numbers or names. This subskill is also assessed in interactive listening, for example, when the test taker needs to understand the context of the upcoming conversation or specific details a professor says about a class assignment. Listening for specific information is also assessed through the dictation task by requiring test takers to demonstrate word-level comprehension. Although the CEFR Manual authors note that dictation is less a communicative-skills task and more meaning-oriented, they do state it may be useful in assessment contexts as a window into linguistic competencies (Council of Europe, 2001, pp. 99–100). For example, dictation assesses orthographic skills as well as phonological, morphosyntactic, and semantic knowledge. Such knowledge is critical in order to be able to extract key words or phrases from a stream of speech. Similarly, listening for a speaker's intent, emotion, and/or attitude involves listening for suprasegmentals like variations in pitch, intonation, or stress patterns in a stream of sound, all of which a test taker must do to accurately complete a dictation. The listener also needs to understand whether a speaker is joyful or irritated, or is emphasizing new or different information, in order to formulate meaning. The longer, turn-based characteristics of speech are directly assessed in the Interactive Listening task because the test taker must select the turn most appropriate for the larger discourse context and relationship between the speakers.

4.2 DET processes

The cognitive processes hypothesized to be activated in listening test tasks are those that would be used in the real world while understanding spoken language. Listening assessment research draws on first language speech science, psycholinguistics, and phonetics, among other fields (for a review of models, see Wagner, 2022). Cognitive validation in second language assessment has been employed to make inferences about listening and reading in particular, which are mental processes that we can only gauge based on external responses such as answers to comprehension questions, eye gazes, or other observable behaviors. Cognitive validity (Glaser, 1991) addresses whether test tasks elicit the internal listening processes that would be used if the test taker were listening in real-world non-test tasks. It is of interest to assessment designers in high-stakes tests where score results are used as a proxy of suitability for a job or a program of study; cognitive processes assessed should map to the cognitive demands of, for instance, listening in university settings (Field, 2013). Cognitive processing may also vary depending on language proficiency level as well as the context for listening and/or the traits of the listener or speaker.

Existing inquiries into listening cognitive processes have focused on not just speech perception and processing but also the question of listener expertise. Beginner listeners must direct their working memory resources to basic processes that, over time and experience, become increasingly automatized. For instance, when listening in an L2, the process of understanding words and meaning from connected speech becomes more automatic, allowing the listener to instead draw on pragmatic and other knowledge when making inferences or judgments, evaluating ideas, and monitoring their own understanding (Field, 2013). This is not to say that context plays no role; listeners must contend with a great deal of speaker variation and have awareness of a variety of speech acts in the real world. In the DET tasks of Dictation and Interactive Listening, the listening skills assessed reflect these cognitive underpinnings.

Dictation has been used as an integrated test of listening and writing even before the advent of recorded media (Bradlow & Bent, 2002, 2008; Buck, 2001; Smith & Kosslyn, 2007). Dictation has received some criticism from language testers, namely that it requires only mimicry or more basic linguistic processes, rather than creative language production. Although dictation does not require overt speech production, it requires test takers to employ their phonological loop, which is better-developed in more able language learners (Baddeley, 1992). Test takers must exercise their processing skills to hold a spoken utterance in working memory long enough to reproduce it verbatim. Language learners whose word recognition skills are still developing will be less able to form meaning from strings of sound (Buck, 2001). Unlike written language on a page or screen, natural, connected speech does not always have word boundaries indicated with pauses (Rost, 2011). The more proficient a test taker is, the better they are at grouping words into phrases or idea units. Those more skilled at this grouping or chunking are able to transcribe longer input audio more accurately and efficiently.

When completing Interactive Listening items, test takers draw on their interactional competence (Galaczi & Taylor, 2018, 2020). Cognitive (Taylor & Geranpayeh, 2011) and sociocognitive (Mislevy, 2018; Weir, 2005) listening elements play roles in this process: test takers must activate not only linguistic knowledge (listening, reading, and writing skills) but also critical thinking skills and academic-setting interactional knowledge, as well as their working memory and executive functioning ability. For example, in an interaction with a student or a professor in the Interactive Listening roleplay, the test taker (listener) must contend with several interrelated processes: using top-down and/or bottom-up processes to understand the individual utterances, all while managing the topic, managing turn taking, and recognizing that the conversation occurs in a particular speech situation or macro-context, refer back to Figure 2. In the A&L framework, listeners may be employing top-down or bottom-up processes, or using metacognitive strategies such as monitoring their own understanding, in both real-world and assessment listening situations.

4.3 DET attributes

Also impacting the nature of listening may be *attributes* which, from the A&L framework, address listener, task, or test-taker characteristics. Task features include (but are not limited to) the recorded media, task format, or expected response (Field, 2013). Task formats must be considered for

not only elicitation of intended cognitive processes, but also accessibility and timing considerations. Certain attribute-related factors pertain to the Dictation and Interactive Listening task types:

- *Input characteristics*: For all audio input, listening stimuli are produced by either voice actors or a text-to-speech (TTS) engine that generates the Duolingo World Character voices. The test uses American English pronunciation (“accent”) common in mass media, education, and commerce. Filled pauses (*uhs*, *ums*), unfilled hesitations, false starts, or repetitions, although commonly found in real speech, are not included in sentence-length stimuli. Speech rate is held constant (averaging three words per second); test takers cannot decrease or increase the speed of the input audio. (See [Naismith et al., forthcoming](#) for a detailed discussion of this attribute.)
- *Timing*: DET items are efficient in that they take a short time to administer while several measurement opportunities are available per item. The measurement information each item provides per unit of time is maximized (for more on item efficiency, see [Crabtree, 2016](#); [Jodoin, 2003](#); [Kim et al., 2022](#); [Wan & Henly, 2012](#)). The task types of relevance to this paper (those which assess listening) are described in Table 1.
- *Other important attributes based on the A&L framework*: We recognize that listener, test, and task-format variables have the potential to play a role in assessment and evaluation. Affective factors such as test-taker or listener anxiety ([Chen, 2012](#)) or motivation ([Ling et al., 2017](#)) are part of this. In addition, interpersonal, intrapersonal, neurological, or experiential factors described in the DET theoretical ecosystem ([Burstein et al., 2022](#)) may play a role in test takers being able to show their listening proficiency. The DET Technical Manual ([Naismith et al., 2025](#)) contains additional information about how the DET prioritizes a delightful* test-taker experience throughout the test, as well as demographic information about the current DET test-taker population.

5 Summary and future directions

At the time of publication of this report, the Duolingo English Test primarily assesses the skill of listening via two task types, Dictation and Integrated Listening. This paper describes these tasks and the underlying listening sub-constructs which they assess, using the Aryadoust and Luo (2023) framework and the CEFR as the theoretical bases. Overall, these two task types assess a wide range of listening subskills and processes and possess a number of important listening attributes. As with all DET task types, these listening tasks are automatically generated and scored with humans in the loop† throughout all processes. Table 1 shows the summary of the task types and target constructs assessed in terms of the subscores on the DET (independent skill subscores and integrated skill subscores) and the A&L approaches (listening subskills, processes, and approaches).

Together, these tasks and assessment processes contribute to the DET mission of improving access to education for all test takers because they allow for efficient and delightful assessment of the listening construct. Because test design is an ongoing and iterative process, assessment of listening on the DET will inevitably continue to evolve. Future directions in listening task development could therefore include research concerning the use of artificial intelligence (AI) to promote further authentic interaction and mediation, as well as including a wider variety or personalized choice of speaker voices.

Table 1. DET Subskills, Processes, and Attributes

Item name	SWRL skills	Integrated skills	Listening subskills	Listening processes	Listening attributes
Listen and Type (Dictation)	Listening, Writing	Conversation, Comprehension, Production	Listening for specific information; listening for detailed understanding; understanding local linguistic meanings	Top-down; bottom-up; memory; cognitive and metacognitive strategy use	Test/task (design; timing; input) and test-taker (affective) factors

*“Delight” is a term from user experience research, used here to refer to maximizing the usability of a product interface to improve test-taker experience.

†Human-in-the-loop AI refers to human–AI system interactions. The DET uses the term to refer to the range of human involvement in building and managing AI systems ([Mosqueira-Rey et al., 2023](#)). Human–system interaction can range from real-time human interaction for improving AI performance ([Wang, 2019](#)) to human oversight at critical decision points, such as human annotation and reviewing efforts ([Association of Test Publishers, 2024](#); [Munro, 2021](#)).

Item name	SWRL skills	Integrated skills	Listening subskills	Listening processes	Listening attributes
Interactive Listening (Interactional competence, see LaFlair et al., 2025)	Directly: Listening, Writing, Reading; Indirectly: Speaking (interaction)	Conversation, Comprehension, Production	Listening for specific information; listening for gist; listening for detailed understanding; listening for implication or inference; communicative listening ability; integrated listening skills	Top-down; bottom-up; memory; cognitive and metacognitive strategy use; managing	Test/task (design; timing; input; genre; task types) and test-taker (affective) factors

6 References

- Alderson, J. C. (2002). *Case studies in the use of the Common European Framework*. Council of Europe.
- Aryadoust, V. (2020). A review of comprehension subskills: A scientometrics perspective. *System*, 88, Article 102180. <https://doi.org/10.1016/j.system.2019.102180>
- Aryadoust, V., & Luo, L. (2023). The typology of second language listening constructs: A systematic review. *Language Testing*, 40(2), 375–409. <https://doi.org/10.1177/02655322221126604>
- Association of Test Publishers. (2024). *Creating responsible and ethical AI policies for assessment organizations*.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
- Baddeley, A. (1992). Working memory. *Science*, 255(5044), 556–559. <https://doi.org/10.1126/science.1736359>
- Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America*, 112(1), 272–284. <https://doi.org/10.1121/1.1487837>
- Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2), 707–729. <https://doi.org/10.1016/j.cognition.2007.04.005>
- Buck, G. (2001). *Assessing listening*. Cambridge University Press.
- 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. <https://go.duolingo.com/ecosystem>
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47. <https://doi.org/10.1093/applin/1.1.1>
- Chen, H. (2012). The moderating effects of item order arranged by difficulty on the relationship between test anxiety and test performance. *Creative Education*, 3(3), 328–333. <https://doi.org/10.4236/ce.2012.33052>
- Council of Europe. (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge University Press.
- Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – companion volume*. Council of Europe Publishing. <https://www.coe.int/lang-cefr>
- Crabtree, A. R. (2016). *Psychometric properties of technology-enhanced item formats: An evaluation of construct validity and technical characteristics* [PhD thesis, University of Iowa]. <https://doi.org/10.17077/etd.922fbj4d>
- Cumming, A. (2014). Assessing integrated skills. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 226–229). John Wiley & Sons, Inc. <https://doi.org/10.1002/9781118411360.wbcla131>
- Field, J. (2008). *Listening in the language classroom*. Cambridge University Press.
- Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (Vol. 35, pp. 77–151). Cambridge University Press.
- Figueras, N., & Noijons, J. (Eds.). (2009). *Linking to the CEFR levels: Research perspectives*. Cito.
- Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, 15(3), 219–236. <https://doi.org/10.1080/15434303.2018.1453816>
- Galaczi, E., & Taylor, L. (2020). Measuring interactional competence. In P. Winke & T. Brunfaut (Eds.), *The routledge handbook of second language acquisition and language testing* (pp. 338–348). Routledge.
- Glaser, R. (1991). Expertise and assessment. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 17–30). Englewood Cliffs.
- Goodwin, S., Attali, Y., LaFlair, G. T., Runge, A., Park, Y., Davier, A. A. von, Yancey, K. P., Naismith, B., & Chuang, P.-L. (2025). *Duolingo English Test: Writing construct* (Duolingo Research Report DRR-22-03). Duolingo. <https://go.duolingo.com/scored-writing>
- Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, 40(1), 1–15. <https://doi.org/10.1111/j.1745-3984.2003.tb01093.x>
- Kim, A. A., Tywoniw, R. L., & Chapman, M. (2022). Technology-enhanced items in grades 1-12 English language proficiency assessments. *Language Assessment Quarterly*, 19(4), 343–367. <https://doi.org/10.1080/15434303.2022.2039659>
- Lado, R. (1961). *Language testing: The construction and use of foreign language tests. A teacher's book*. Longmans, Green and Company.
- Lado, R. (1964). *Language teaching, a scientific approach*. McGraw-Hill.
- LaFlair, G. T. (2020). *Duolingo English Test: Subscores* (DRR-20-03). Duolingo. <https://duolingo-papers.s3.amazonaws.com/reports/subscore-whitepaper.pdf>
- LaFlair, G. T., Runge, A., Attali, Y., Park, Y., Church, J., & Goodwin, S. (2025). *Interactive Listening—The Duolingo English Test* (Duolingo Research Report DRR-23-01). <https://go.duolingo.com/interactive-listening-whitepaper>
- Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test? *Applied Psychological Measurement*, 41(7), 495–511. <https://doi.org/10.1177/0146621617707556>
- Martyniuk, W. (Ed.). (2010). *Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's draft manual*. Cambridge University Press.
- McNamara, T. (2000). *Language testing*. Oxford University Press.
- Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

- Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, A. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56, 3005–3054. <https://doi.org/10.1007/s10462-022-10246-w>
- Munro, R. (2021). *Human-in-the-loop machine learning: Active learning and annotation for human-centered AI*. Manning Publications.
- Naismith, B., Cardwell, R., LaFlair, G., Nydick, S., & Kostromitina, M. (2025). *Duolingo English Test: Technical manual* [Duolingo Research Report]. Duolingo. <https://go.duolingo.com/dettechnicalmanual>
- Naismith, B., Goodwin, S., & Kang, O. (forthcoming). Speech rate in L2 listening assessment. In E. Wagner & Y. In'nami (Eds.), *The theory and practice of second language listening assessment*. Applied Linguistics Press.
- Ockey, G., & Wagner, E. (2018). *Assessing L2 listening: Moving towards authenticity*. John Benjamins. <https://doi.org/10.1075/llt.50>
- Oxford, R. (2002). *Integrated skills in the ESL/EFL classroom*. ERIC Digest [Electronic resource]. <http://eric.ed.gov/?id=ED456670>
- Park, Y., Cardwell, R., Goodwin, S., Naismith, B., LaFlair, G. T., Loh, Kai-Ling, Yancey, Kevin P., & Zhang, X. (2025). *Assessing speaking on the Duolingo English Test* (Duolingo Research Report DRR-23-03). <https://go.duolingo.com/listening>
- Park, Y., LaFlair, G. T., Attali, Y., Runge, A., & Goodwin, S. (2022). *Duolingo English Test: Interactive reading* (DRR-22-02). Duolingo. <https://go.duolingo.com/interactive-reading>
- Richards, J. C. (2008). *Teaching listening and speaking: From theory to practice*. Cambridge University Press.
- Rost, M. (2011). *Teaching and researching listening* (2nd ed.). Pearson.
- Smith, E. E., & Kosslyn, S. M. (2007). *Cognitive psychology: Mind and brain*. Pearson Prentice Hall.
- Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalising the test construct. *Journal of English for Academic Purposes*, 10(2), 89–101. <https://doi.org/10.1016/j.jeap.2011.03.002>
- Wagner, E. (2022). L2 listening comprehension: Theory and research. In E. H. Jeon & Y. In'nami (Eds.), *Understanding L2 proficiency: Theoretical and meta-analytic investigations* (pp. 213–233). John Benjamins.
- Wan, L., & Henly, G. A. (2012). Measurement properties of two innovative item formats in a computer-based test. *Applied Measurement in Education*, 25(1), 58–78. <https://doi.org/10.1080/08957347.2012.635507>
- Wang, G. (2019). *Humans in the loop: The design of interactive AI systems*. <https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems>
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave MacMillan.