Assessing Listening on the Duolingo English Test

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., 2022 for writing; LaFlair et al., 2023 for interactional competence; Park et al., 2023 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 (2023). 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 item types.

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

∗ Duolingo, Inc.

Corresponding author:
Sarah Goodwin
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 The Duolingo English Test ........................................ 3
   2.1 Dictation item type ........................................... 4
   2.2 Interactive listening item type ............................. 5

3 Theoretical background ........................................... 6
   3.1 General theoretical background ............................ 6
   3.2 Listening theoretical background .......................... 7

4 Listening approaches and the DET ............................. 8
   4.1 DET subskills .................................................. 8
   4.2 DET processes ................................................. 10
   4.3 DET attributes ................................................ 11

5 Summary and future directions ................................. 12

6 References .................................................................. 14
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 (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., 2023 for writing; LaFlair et al., 2023 for interactional competence; Park et al., 2023 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 item 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 integrated skills subscores (LaFlair, 2020): Literacy (reading and listening), Comprehension (reading and listening), Conversation (speaking and listening), and Production (speaking and writing). Item types which assess listening therefore contribute to the Comprehension and Conversation subscores. (See Cardwell et al., 2023, 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., 2023 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).

In total, there are 13 item types on the DET, all of which are scored. In the first phase of the test, there are four item types which are part of the computer-adaptive administration: yes / no vocabulary, C-test, elicited imitation, and dictation. These item 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 eight more authentic task types comprising interactive reading, interactive listening, picture descriptions (speaking and writing), and extended production (speaking and writing). The two item types that assess listening directly are dictation and interactive listening.

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.
2.2 Interactive listening item type

For interactive listening, test takers participate in a simulated conversation by choosing responses from multiple-choice options. They then write a short summary of the conversation they just had. In the interaction, the test taker plays the role of a university student, conversing either with another student or a professor. They first read a scenario that sets the scene for the interaction. Next, they click a button with a play icon to play each utterance; 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. 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. (2023).
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).*

*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.
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 recent 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 (Figure 3).

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, 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 test items 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 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).

Reception and Interaction from the CEFR both pertain to listening comprehension ability. The CEFR distinguishes language reception from interaction, treating language reception in isolation as people employing reading or listening skills in relative isolation (e.g., listening to announcements on public transportation). Reception is therefore the primary CEFR component assessed by the DET dictation task. With reception as a process being required for interaction, interaction is conceptualized as more than just receiving or producing utterances, for example,
talking turns, cooperating, and seeking clarification (Council of Europe, 2001, p. 14). Interaction is therefore a primary focus of the DET interactive listening task.†

The necessity for 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 speakers’ intention or rhetorical purpose—how or why someone said something. The DET listening tasks attempt to cover this spectrum of purposes; while the dictation task focuses more on the details of listening input, the interactive listening task focuses on elements such as speaker intention. Take for example the earlier illustrations of these two tasks (Figures 1 and 2). For dictation, 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 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 specific details a professor says about a class assignment or advising. 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

† In addition, listening is assessed as a secondary construct as part of the DET speaking-focused task Listen, Then Speak. See Park et al. (2023) for further detail about these tasks.
directly assessed in the DET 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 DET 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 skills 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 [Galaczi and Taylor (2018), p. 227; refer back to Figure 2 for an example]. 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, item 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 items:

- **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 (uhhs, umms), 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.

- **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 item 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 (Cardwell et al., 2023) 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.

---

1“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.

© 2023 Duolingo, Inc
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 item 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 item types assess a wide range of listening subskills and processes and possess a number of important listening attributes. As with all DET item types, these listening tasks are automatically generated and scored with humans in the loop throughout all processes. Table 1 shows the summary of the item types and target constructs assessed in terms of SWRL skills, subscores, and the A&L 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 item 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>
<thead>
<tr>
<th>Item name</th>
<th>SWRL skills</th>
<th>Integrated skills</th>
<th>Listening subskills</th>
<th>Listening processes</th>
<th>Listening attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listen and Type (Dictation)</td>
<td>Listening, Writing</td>
<td>Conversation, Comprehension, Production</td>
<td>Listening for specific information; listening for detailed understanding; understanding local linguistic meanings</td>
<td>Top-down; bottom-up; memory; cognitive and metacognitive strategy use</td>
<td>Test/task (design; timing; input) and test-taker (affective) factors</td>
</tr>
<tr>
<td>Interactive Listening</td>
<td>Directly: Listening, Writing</td>
<td>Conversation, Comprehension, Production</td>
<td>Listening for specific information; listening for gist; listening for detailed understanding; listening for implication or inference; communicative listening ability; integrated listening skills</td>
<td>Top-down; bottom-up; memory; cognitive and metacognitive strategy use; managing</td>
<td>Test/task (design; timing; input; genre; task types) and test-taker (affective) factors</td>
</tr>
</tbody>
</table>
6 References


© 2023 Duolingo, Inc


Rost, M. (2011). *Teaching and researching listening* (2nd ed.). Pearson.


