Duolingo English Test: Technical Manual

Duolingo Research Report
July 9, 2025 (49 pages)
https://englishtest.duolingo.com/research

This document **is not** the most recent version of the Duolingo English Test Technical Manual and may not represent the most recent version of the test. For the most recent technical manual, please visit https://go.duolingo.com/dettechnicalmanual.

Ben Naismith*, **Ramsey Cardwell***, **Geoffrey T. LaFlair***, **Steven Nydick***, **and Masha Kostromitina***

Abstract

The Duolingo English Test Technical Manual provides an overview of the design, development, administration, and scoring of the Duolingo English Test. Furthermore, the Technical Manual reports validity, reliability, and fairness evidence, as well as test-taker demographics and the statistical characteristics of the test. This is a living document whose purpose is to provide up-to-date information about the Duolingo English Test, and it is updated on a regular basis.

Last Update: July 9, 2025

Contents

Note: We would like to acknowledge the contributions of Burr Settles, the creator of the Duolingo English Test and author of the first Technical Manual, and Alina von Davier, for her influence on the design and strategy of this manual.

*Duolingo, Inc.

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

1 Introduction

The Duolingo English Test (DET) is a measure of English language proficiency for communication and use in English-medium settings, covering the full range of language proficiency. It assesses test-taker ability to use the language skills of speaking, writing, reading, and listening (SWRL), as well as the integration of these skills for literacy, conversation, comprehension, and production. The test is designed for maximum accessibility while maintaining high measurement accuracy and using authentic multimodal inputs; it is delivered via the internet, without a testing center, and is available 24 hours a day, 365 days a year. In addition, as a computer-adaptive test (CAT), it is designed to be efficient; the test takes approximately one hour to complete, though as a CAT the exact time varies for each test taker. The test uses a variety of task types that provide maximal coverage of the English language proficiency construct while being feasible to develop, administer, and score at scale. In all areas of the test, high standards of security and psychometric quality are maintained.

In adhering to the *Standards for Educational and Psychological Testing* (Standards; AERA et al., 2014) as they relate to test documentation (Chapter 7), this technical manual provides an overview of all aspects of the DET so that stakeholders can make informed decisions about how to interpret and use DET test scores. Like the *Standards* themselves, this technical manual begins with more theoretical foundations before proceeding to operational topics. It contains a presentation of:

- the test's tasks and the constructs they cover
- how the test is developed using human-in-the-loop* generative AI
- the adaptive delivery and scoring of test items using computational psychometrics
- the maintenance of score comparability and reliability
- the relationships of test scores with other variables (e.g., scores on other tests)
- considerations of fairness and beneficial consequences of test use
- the multi-layered approach to test security combining automated detection and expert human review

Collectively these topics cover the different types of validity evidence described in the *Standards* and the inferences in an argument-based validation framework (e.g., Chapelle, 2021). Argument-based test validation frameworks are an approach to systematically evaluating the validity of a test by constructing and analyzing a structured argument grounded in evidence and theoretical reasoning, to assess whether there is sufficient support for the test's intended interpretations and uses (Kane, 2012).

Since its inception in 2016, the social mission of the DET has been to lower barriers to education access for English language learners around the world. The DET achieves this goal by leveraging technological advances in annual test updates to produce an accessible and affordable high-stakes language proficiency test that produces valid, fair, and reliable test scores. These scores are intended to be interpreted as reflecting test-taker English language proficiency and to be used in a variety of settings, including for post-secondary admissions decisions. To date, the success of this mission is evidenced by the widespread adoption of the DET by over 5,900 academic programs in more than 100 countries.

At the end of each section of the technical manual, we provide links to further readings that expand on the topics discussed in the section. The 📖 icon indicates a peer-reviewed publication (e.g., a journal article or conference proceedings), the 📋 icon represents an internal DET publication (e.g., a technical report or white paper), and the icon represents a DET blog post.

Further readings

- The evolution of the Duolingo English Test
- How many universities accept the Duolingo English Test
- Why standardized tests need to embrace AI

*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).

2 Theoretical Basis

The Duolingo English Test employs a novel assessment ecosystem (Burstein et al., 2022; Langenfeld et al., 2022) composed of an integrated set of frameworks related to language assessment, design, psychometrics, test security, and test-taker experience (TTX). Furthermore, the processes and documentation of the DET—including test development, scoring, and documentation of validity, reliability and fairness evidence—have been externally evaluated against the AERA et al. (2014) standards and are continually internally evaluated against the Responsible Artificial Intelligence (RAI) Standards (Burstein, 2025). These theoretical underpinnings motivate the research philosophy and values of the DET, which aim to make the DET test-taker-centered by taking advantage of the latest developments in technology (including artificial intelligence[*]), applied linguistics, psychometrics, and assessment science.

On a more fundamental level, the DET subscribes to the interactionalist perspective of what a test can in fact assess (Chapelle, 1998; Messick, 1989, 1996; Young, 2011). In this conceptualization, test-taker performance reflects two elements and their interaction: 1) the underlying traits of the test taker (English language skills, strategies, and competencies), and 2) the context-specific behaviors of the test taker (task performance). For example, an individual may evidence a certain level of spoken proficiency during a face-to-face conversation yet struggle with the exact same conversation on the phone. It is therefore necessary to always consider the characteristics of the setting (including the task type and language modality) when drawing conclusions about a test taker's underlying traits. This theory of language aligns closely with the tenets of the communicative language ability (CLA) model, which calls for language assessments to be informed by "language ability in its totality" (Bachman & Palmer, 1996, p. 67). Such an approach is also consistent with the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001, 2020), which conceptualizes language use as the leveraging of communicative language competences in various contexts with corresponding conditions and constraints (p. 32).

The result of the above considerations is a modern test that equally meets the assessment criteria and the needs of stakeholders, and which is continually being evaluated and iterated upon in all aspects of our assessment processes. Together, these ecosystem frameworks, testing standards, and research philosophy support a test validity argument built on a digitally informed chain of inferences, appropriate for a digital-first assessment of this nature and consistent with professional standards of practice. As a result, the adaptive DET can be seen to assess test takers' proficiency in General English and English for Academic Purposes (EAP), both of which are essential for success in a range of academic or professional settings. The DET considers EAP to consist of the language knowledge and skills necessary to perform common communicative and pedagogical tasks in a range of educational contexts across academic disciplines. This definition highlights the importance of communicative competence in educational settings.

2.1 Test Constructs

Here we describe the constructs being tested, that is, "the specific definition of an ability that provides the basis for a given assessment or assessment task and for interpreting scores derived from this task" (Bachman & Palmer, 2010, p. 43). The DET measures test-taker ability to use the independent language skills of speaking, writing, reading, and listening (SWRL skills). These subskills can also be combined into the integrated language skills required for literacy (reading and writing), conversation (speaking and listening), comprehension (reading and listening), and production (speaking and writing), including the skills necessary for success in academic contexts. These independent and integrated skill areas correspond to the eight DET subscores. (For white papers on how the DET assesses each SWRL skill, see Park et al. (2023) for speaking; Goodwin et al. (2023) for writing; Park et al. (2022) for reading, and Goodwin and Naismith (2023) for listening.) In addition, certain DET task types target the assessment of vocabulary[†] because vocabulary knowledge is critical to proficiency in all language skills areas. (See Park et al. (2024) for how the DET assesses vocabulary.) Figure 1 shows the relationship between the SWRL constructs/subscores, integrated skills constructs/subscores, and DET overall score.

[*] Defined here as "technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity and autonomy" (Stryker & Kavlakoglu, 2024).

[†] Here we use the term "vocabulary" in the broad sense, also referred to as "lexis", which includes the knowledge and use of words, word parts, multi-word units, and the connections between them.

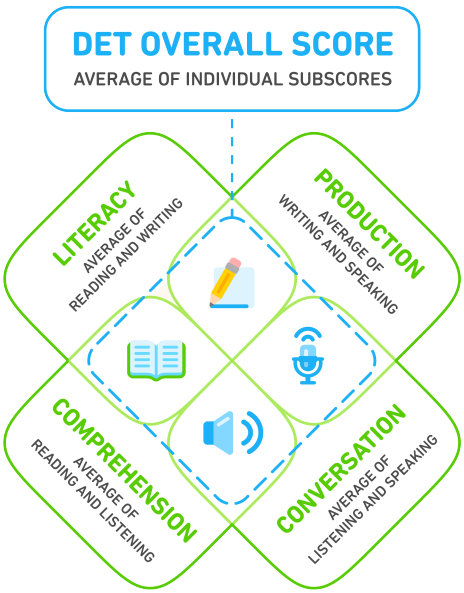

Figure 1. Relationship between SWRL subscores, integrated subscores, and DET overall score

In total, the DET has 13 different graded task types that collectively measure test-taker proficiency in the English-language constructs described above. The creation and selection of this specific combination of task types is guided by the DET ecosystem (Burstein et al., 2022), especially the Language Assessment Design Framework. In this framework, task design and scoring target constructs relevant for General and Academic English language proficiency. Test security is also an aspect of the DET ecosystem, and having a variety of task types provides one layer of protection against certain cheating strategies. In addition to test use validity and security, another consideration in test design is ensuring a delightful TTX. As a result of these considerations, DET tasks are intuitive, reducing the need for test-specific preparation. All DET task types are summarized in Tables 1 and 2 and are described individually in the subsequent section.*

Further readings

- A theoretical assessment ecosystem for a digital-first assessment
- The Duolingo English Test Responsible AI standards
- How does the DET measure writing?
- Assessing Speaking on the Duolingo English Test
- Assessing Listening on the Duolingo English Test

*See Section 5 for information on subscores.

Table 1. Constructs and Task Types

Individual skill	Description	Integrated skill contribution	Task types
Speaking	Producing spoken English from basic discourse to advanced discourse at CEFR levels A1–C2	Conversation Production	• Extended Speaking • Interactive Speaking • Picture Description (speaking) • Speaking Sample
Writing	Writing English texts with a variety of rhetorical functions at CEFR levels A1–C2	Literacy Production	• Interactive Writing • Interactive Listening (summarization) • Picture Description (writing) • Writing Sample
Reading	Comprehending written English from basic informational texts to advanced expository/persuasive texts at CEFR levels A1–C2	Literacy Comprehension	• Interactive Reading • C-test • Vocabulary in Context • Yes/No Vocabulary
Listening	Comprehending spoken English from basic discourse to advanced discourse at CEFR levels A1–C2	Conversation Comprehension	• Interactive Listening • Dictation

3 Test Task Types

We now describe the 13 task types (and their sub-tasks) from Table 1. These task specifications and descriptions of the item generation processes (see Section 4) constitute part of the test's content-oriented validity evidence.The DET task types include both closed-ended task types (e.g., C-test and Yes/No Vocabulary) and open-ended task types (e.g., Picture Description and Writing Sample). Many of these tasks are integrative, requiring the test taker to demonstrate proficiency with multiple skills, for example Dictation (listening and writing) or C-test (reading and writing). Some task types are also multimodal, incorporating images, animations, audio, and written text.

Tasks vary too in their authenticity, that is, "the degree of correspondence of characteristics of a given language test task to the features of a TLU [target language use] task" (Bachman & Palmer, 2010, p. 23). For example, highlighting relevant text in a reading passage as part of the Interactive Reading task is highly authentic as it is a skill that many test takers will employ in academic (or other) contexts. In contrast, the Yes/No Vocabulary task, which targets vocabulary knowledge, is less authentic, as deciding if a word is real or not is not an activity that test takers are likely to encounter outside of a language testing context. The DET deliberately includes tasks with different levels of authenticity in order to maximize the efficiency of the test for measurement accuracy while also fully covering the intended constructs.

Each test task type corresponds to one individual skill construct and two integrated skill constructs. We now describe each task in turn, organized by individual skill in order of SWRL. Within each subsection, we begin with the more authentic tasks. For the order in which tasks are presented on the test and the number of each task type administered in a test session, see Section 5.1.

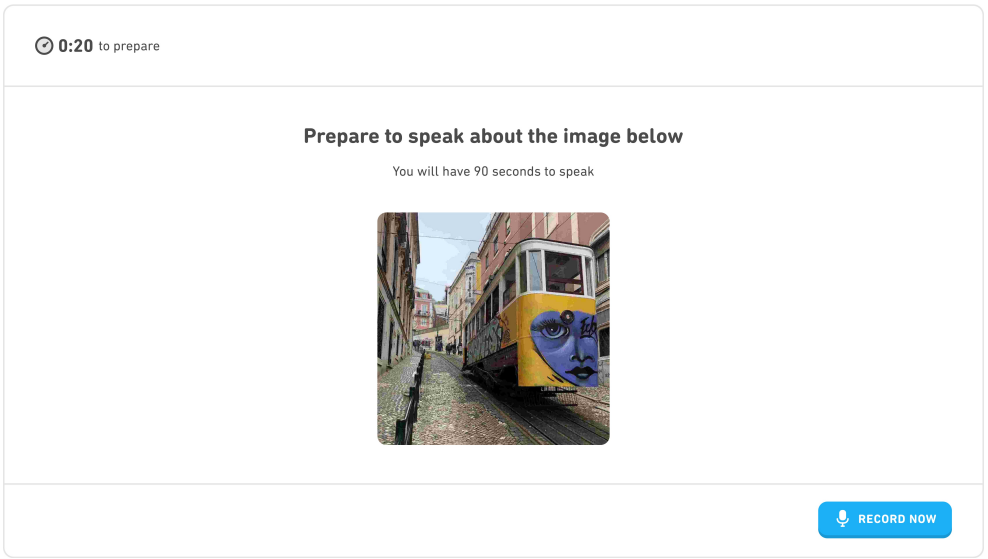

Figure 2. Example Picture Description (speaking) Task

3.1 Speaking Tasks

Each test session includes multiple task types requiring an open-ended spoken response. These speaking tasks are administered after the computer-adaptive portion of the test. In this subsection we describe the speaking task types Picture Description, Extended Speaking, Speaking Sample, and Interactive Speaking. These speaking tasks require test takers to respond by speaking in response to various stimuli (text prompts, audio prompts, and images). All of these task types require test takers to speak for an extended time period and to leverage different aspects of their organizational knowledge (e.g., grammar, vocabulary, and discourse coherence) and functional elements of their pragmatic language knowledge (e.g., ideational knowledge; Bachman & Palmer, 1996). All speaking task types elicit responses that evidence proficiency in terms of the speaking subconstructs of Content, Discourse coherence, Grammar, Vocabulary, Fluency, and Pronunciation. Topics are discussed in the different domains described in the CEFR (Personal, Public, Educational, and Professional).

Picture Description

Picture description tasks are a well-established tool for eliciting longer monologic responses, with the contents of the images providing ample input (Boers, 2018; N. de Jong & Vercellotti, 2016; Koizumi & In'nami, 2024). Such prompts are nevertheless sufficiently open-ended to allow for a wide range of possible responses to demonstrate speaking proficiency (Rossiter et al., 2008). Their rhetorical purpose of "description" is essential for effectively communicating in a range of authentic contexts (Qiu, 2022). Picture description tasks are also widely used in L2 assessment contexts, for both speaking and writing, because they do not require written or aural prompts which may impede task comprehension for some, especially lower-proficiency, test takers (Boers, 2018).

In the DET Picture Description (speaking) task type, test takers have 90 seconds to describe a single photograph. The images contain stimulating depictions of people, animals, and objects in a wide range of contexts, enabling test takers of all ability levels to demonstrate lexical and grammatical proficiency (see Figure 2). The stimuli (i.e., the photos) were selected by people with graduate degrees in applied linguistics.

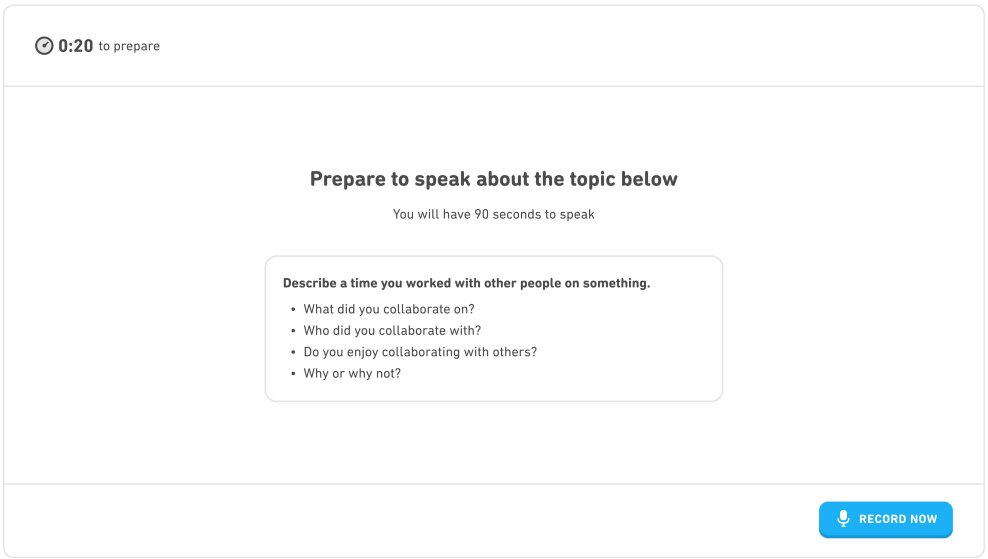

Figure 3. Example Extended Speaking

Extended Speaking and Speaking Sample

Extended, monologic, open-ended speaking tasks are a long-standing staple of language proficiency assessments. These types of oral presentations permit for a wide variety of criteria to be assessed (O'Sullivan, 2008), as previously described, as they evidence both linguistic and strategic competencies (N. H. de Jong, 2023) including discourse competence (Iwashita & Vasquez, 2015). For example, proficiency in the use of numerous informational discourse functions can be elicited, such as providing personal and non-personal information, elaborating, expressing and justifying opinions, and expressing preferences (O'Sullivan, 2008).

For the DET, The Extended Speaking and Speaking Sample tasks require test takers to respond to a single written prompt in either 90 seconds (Extended Speaking; see Figure 3) or three minutes (Speaking Sample; see Figure 4). These prompts ask test takers to recount an experience (i.e., narrative rhetorical function) or argue a point of view (i.e., argumentative/persuasive rhetorical function). A recording of a test taker's spoken response to the Speaking Sample task is provided to institutions with which the test taker shares their results.

Interactive Speaking

The Interactive Speaking task type complements the monologic, single-turn tasks described above by simulating communicative exchanges. To do so, it introduces adaptively sequenced prompts delivered in a conversational format. The task is designed to engage test takers in informal, topic-focused conversation that reflects real-world discourse functions such as narrating, describing, explaining, and expressing opinions (O'Sullivan, 2008). As with the Extended Speaking and Speaking Sample tasks, the construct underlying Interactive Speaking is informed by a multi-dimensional model of speaking proficiency that includes both linguistic and strategic competence; here, however, there is additional emphasis on real-time processing abilities such as rapid planning and speech-act prediction (N. H. de Jong, 2023). While test takers do not themselves initiate turns or ask questions, their responses occur within a structured series of topic-based question–answer adjacency pairs, a common interactional organization in speaking assessments that enables controlled elicitation of dynamic speech (Seedhouse & Harris, 2011). By including this task format in the DET portfolio of speaking tasks, components of speaking proficiency related to communicative competence are better represented, such as the ability to adapt responses to an interlocutor. In addition, interactive speaking task formats such as interviews have been shown to elicit distinct language features from other types of speaking tasks such as monologues (Ahmadi & Sadeghi, 2016).

Each Interactive Speaking task (see Figure 5) consists of a single item comprising a short conversation centered around several topics (usually two). For each topic, test takers hear a series of prompts (usually three or four) delivered by an on-screen avatar. After hearing each prompt, test takers have 35 seconds to record their spoken response. Follow-up questions are selected

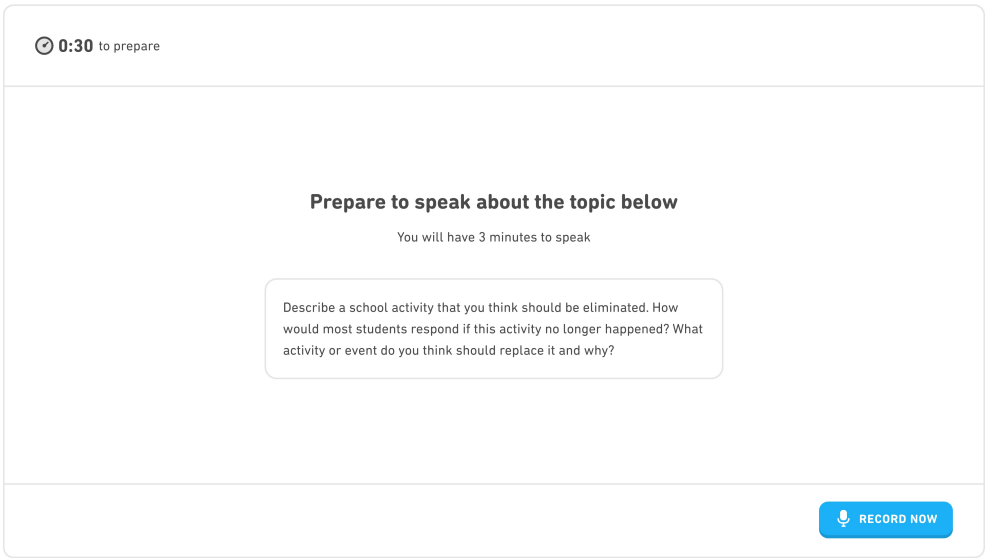

Figure 4. Example Speaking Sample Task

dynamically based on how well test takers addressed previous prompts and what they said, thereby avoiding redundancy and adjusting to the test taker's estimated proficiency. This adaptive delivery is designed to simulate a natural progression within each conversation while ensuring broad topical coverage across test takers. Visual and textual cues help orient the test taker throughout, including clear instructions and automated transitions between prompts and responses.

3.2 Writing Tasks

In this subsection we describe the three open-ended writing task types that measure test takers' English writing abilities: Picture Description (writing), Interactive Writing, and Writing Sample.* All writing task types elicit written responses that evidence writing proficiency in terms of the writing subconstructs of Content, Discourse coherence, Grammar, and Vocabulary. As with the speaking tasks described previously, test takers must demonstrate proficiency in discussing topics in the different domains described in the CEFR (Personal, Public, Educational, and Professional).

Picture Description (writing)

Picture description tasks provide opportunities for test takers to produce written descriptions, which is an important skill necessary in many contexts, including both general and academic (Barkaoui, 2024; Coker, 2012). Although seemingly simple, written picture description requires the use of numerous skills and linguistic structures in order to convey information about the image, to establish a perspective or context in which to situate the objects, and to organize the text so as to focus on salient features of the image (Schleppegrell, 1998).

In the DET Picture Description (writing) task, test takers have 60 seconds to describe a single photograph in writing. As with the Picture Description (speaking) task, the stimuli (i.e., the photos) were selected by people with graduate degrees in applied linguistics. These images are designed to give test takers the opportunity to display their full range of written language abilities as they contain stimulating depictions of people, animals, and objects in a wide range of contexts. The images are therefore capable of eliciting writing from test takers across the proficiency spectrum.

*In addition, there is a written summarization sub-task as part of the Interactive Listening task. This task type is described in the Listening Tasks section.

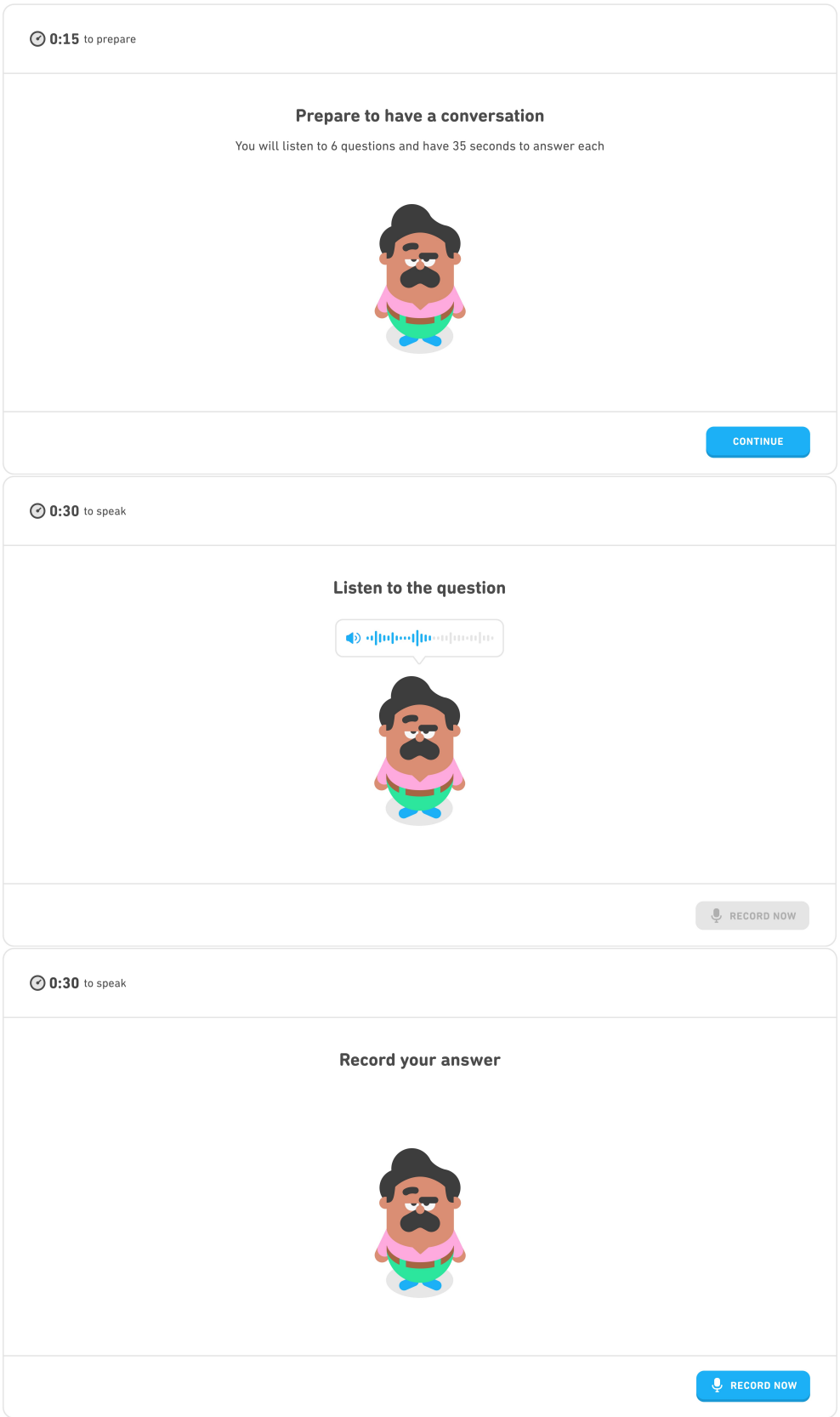

Figure 5. Example Interactive Speaking Task

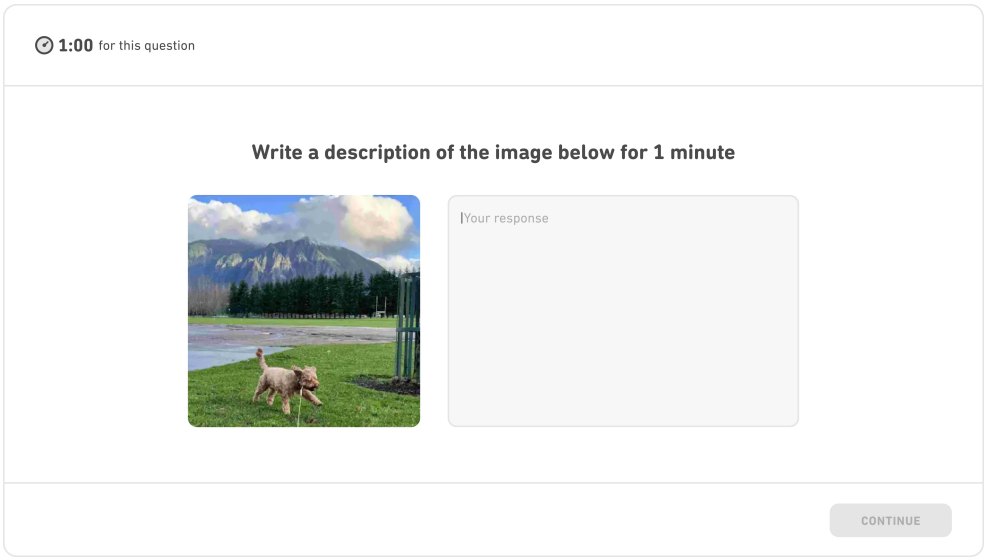

Figure 6. Example Picture Description (writing) Task

Interactive Writing and Writing Sample

The Interactive Writing and Writing Sample tasks are considered independent writing tasks, requiring test takers to demonstrate more discursive knowledge of writing in addition to language knowledge (Weigle, 2002). Independent writing tasks aim to assess test takers' latent abilities to produce writing based on their own ideas, experiences, and knowledge, without the support of external sources (Crossley et al., 2014; Guo et al., 2013). Tasks of this nature are commonly viewed as complex performance activities that reflect authentic writing practices, making them appropriate for language assessments (Enright & Quinlan, 2010). Research has shown that independent writing tasks engage key writing processes—such as planning, monitoring, and revising—that differ by proficiency level (Barkaoui, 2011). Direct assessments of writing in formats such as these are also associated with strong construct, content, and face validity (Hamp-Lyons, 1990).

For both Writing Sample and Interactive Writing, the written prompts ask test takers to recount an experience or argue a point of view, with each rhetorical function (narrative or persuasive) requiring the demonstration of different linguistic resources, discourse features, etc. to successfully achieve the task. For Writing Sample, a single written prompt is presented to the test taker who has up to five minutes to respond. The Writing Sample is the final independent writing task in a test administration (see Figure 8); a test taker's written response to this task is provided to institutions with which the test taker shares their results.

In the case of Interactive Writing (see Figure 7), there are two parts. First, test takers are asked to respond to a prompt (as in Writing Sample) for five minutes. Next, their first response is analyzed in real-time for topic-relevant themes, and they are asked a related follow-up question based on their initial response; they then have three additional minutes to respond. This design is intended to more authentically reflect the real-world writing scenarios and to elicit greater evidence of test takers' true writing proficiency by prompting elaboration.

3.3 Reading Tasks

In this subsection we describe the test task types Interactive Reading, C-test, Vocabulary in Context, and Yes/No Vocabulary. These tasks primarily involve reading, although C-test and Vocabulary in Context require test takers to respond with written input.

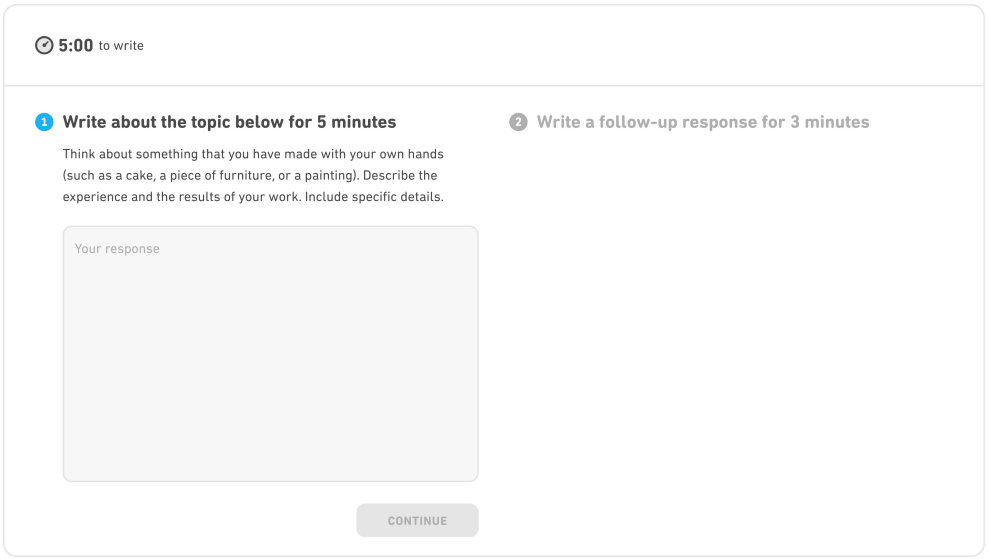

Figure 7. Example Interactive Writing Task

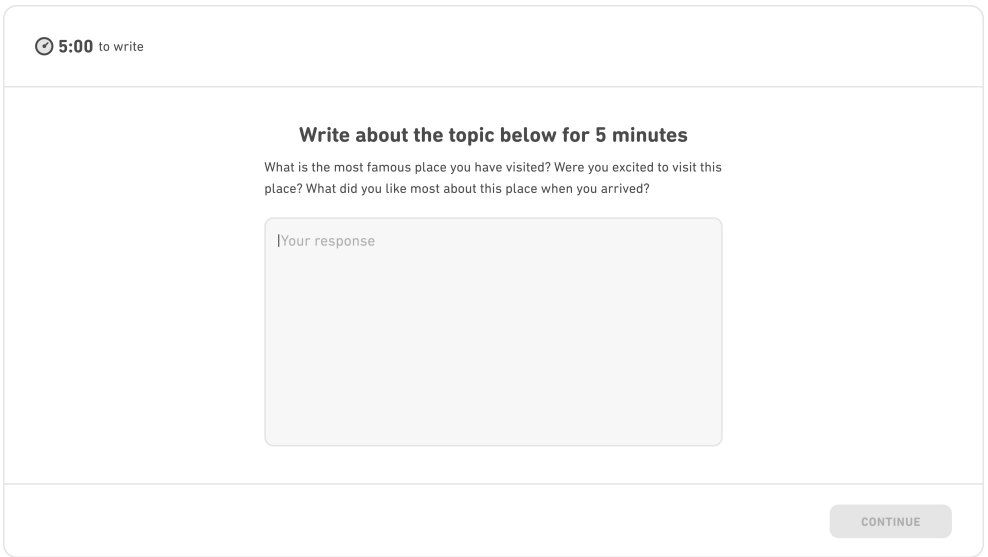

Figure 8. Example Writing Sample Task

Interactive Reading

The Interactive Reading task type (Park et al., 2022) focuses on reading comprehension, complementing the other reading task types (C-test and Yes/No Vocabulary; see next sections) that have a greater focus on bottom-up reading processes (e.g., decoding) and language knowledge. Interactive Reading requires test takers to engage with a text by sequentially performing a series of sub-tasks tapping different subconstructs of reading (reading for specific information, reading for detail, reading for orientation, identifying cues and inferring) and all using the same text as the stimulus. This task type is interactive in the sense that as the test progresses, different portions of the text are presented to the test taker, ensuring a comprehensive evaluation of their reading skills. By incorporating interactivity, the DET interactive tasks provide a dynamic and immersive testing experience that better reflects real-world language usage.

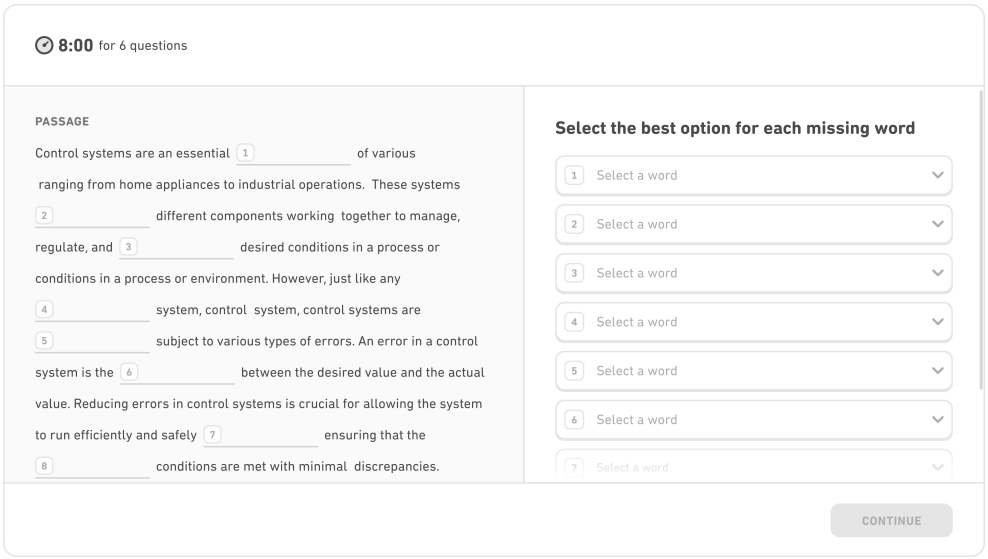

Figure 9. Example Interactive Reading "Complete the Sentences" Sub-task

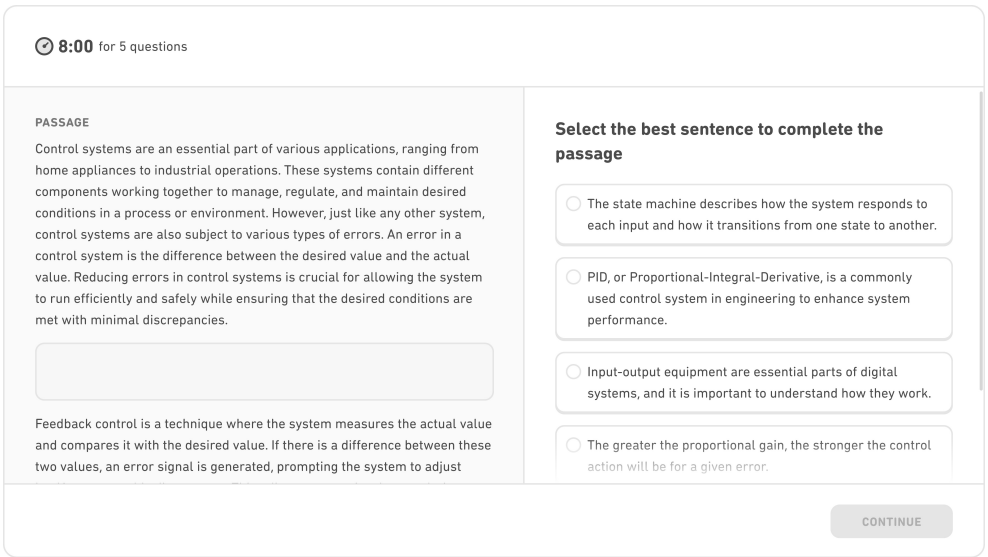

Figure 10. Example Interactive Reading "Complete the Passage" Sub-task

The first sub-task shows the test taker the first half of the text with 5–10 words missing (see Figure 9); test takers must select the word that best fits each blank. Next, test takers are shown the remainder of the text with one sentence missing (see Figure 10); test takers must select the sentence that best completes the passage from among several options. With the text now complete, test takers are shown sequentially two questions and asked to highlight the part of the text that contains the answer (see Figure 11). Test takers are then asked to select an idea that appears in the passage from among several options, only one of which is correct (see Figure 12). Finally, test takers are asked to choose the best title for the text from among several options (see Figure 13).

Each Interactive Reading passage is classified by genre as either narrative or expository; each test taker receives one narrative passage and one expository passage. Passages cover a range of topics and reflect the educational and occupational domains of language use. Additionally, the number of complete-the-sentence blanks across the two tasks is controlled such that each test taker receives approximately the same number. Consequently, in each test session, one Interactive Reading item takes seven minutes, and one Interactive Reading item takes eight minutes.

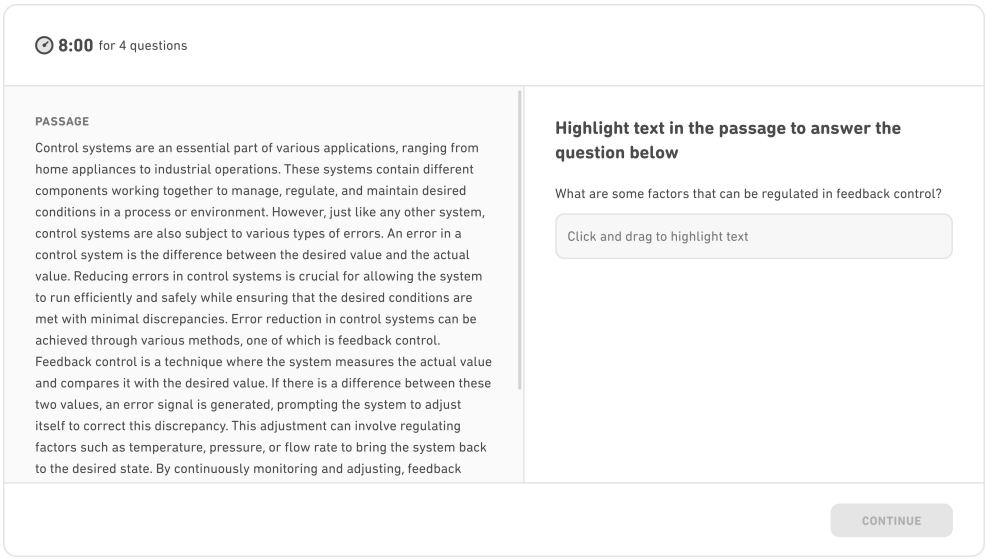

Figure 11. Example Interactive Reading "Highlight the Answer" Sub-task

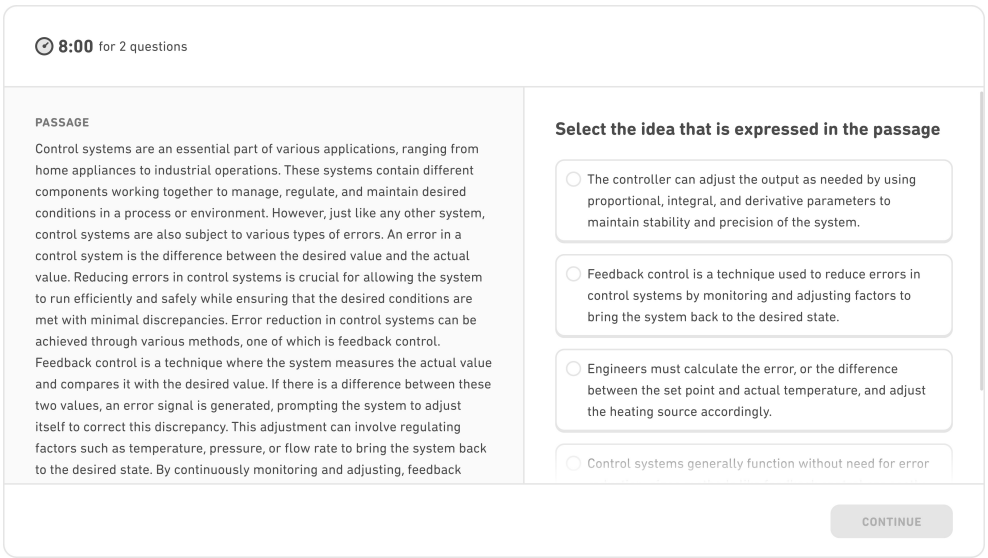

Figure 12. Example Interactive Reading "Identify the Idea" Sub-task

C-test

The C-test task type (see Figure 14) measures a test taker's global language proficiency in the written modality (Norris, 2018), capturing chiefly knowledge of vocabulary and grammar (Eckes & Grotjahn, 2006). In addition, C-test scores correlate moderately well with discrete language components including reading ability (Khodadady, 2014; Klein-Braley, 1997), spelling skills (Khodadady, 2014), and vocabulary (Karimi, 2011). It has been shown that scores from C-tests are significantly correlated with scores from many other major language proficiency tests (Daller et al., 2021; Khodadady, 2014).

In this task, the test taker is presented with a short text. The first and last sentences of the text are fully intact, while alternating words in the intervening sentences are "damaged" by deleting the second half of the word. Test takers respond to the C-test items by completing the damaged words in the passage. Test takers need to rely on context and discourse information to

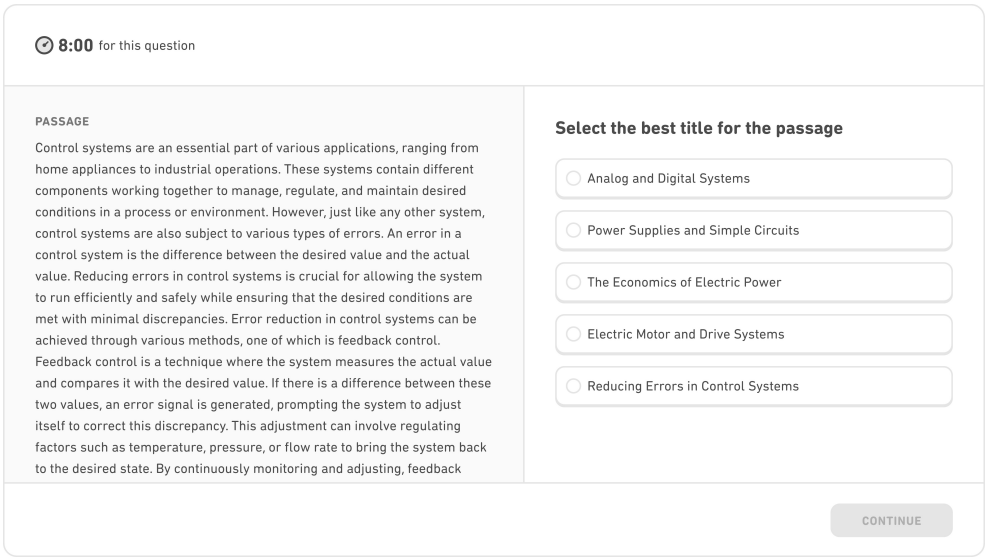

Figure 13. Example Interactive Reading "Title the Passage" Sub-task

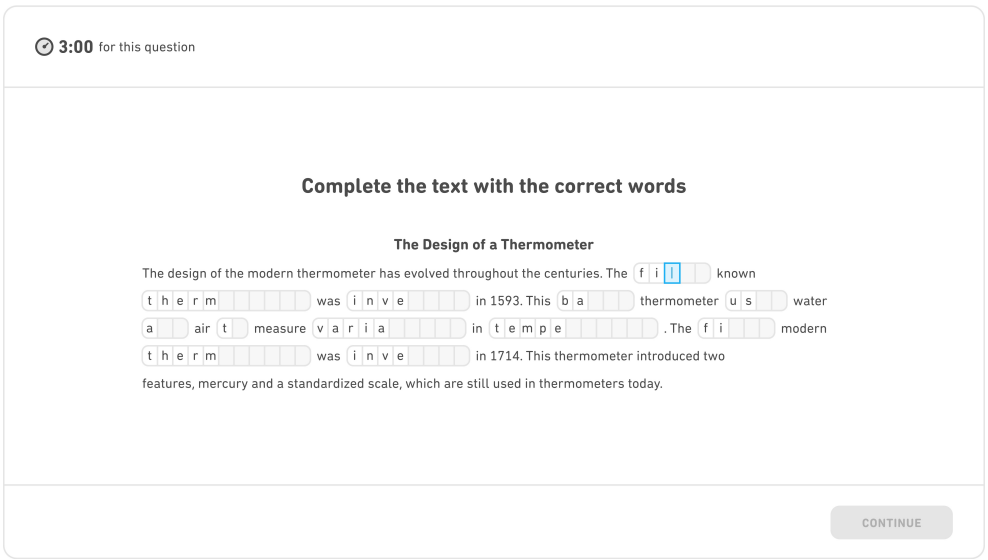

Figure 14. Example C-test Task

reconstruct the damaged words, which include both function and content words spanning multiple parts of speech (i.e., lexical and morphosyntactic categories). The task thus relates to linguistic subskills including reading for information, inferring, and orthographic control (i.e., adherence to spelling and punctuation conventions).

The C-test passages themselves reflect a range of different authentic text types from the educational, professional, public, and private domains of language use, including fiction (e.g., colloquial narratives), news articles, and textbook passages. The linguistic features of these passages have been carefully analyzed to ensure a variety of text types and difficulty levels. In total, more than 150 linguistic features are annotated and accounted for, including features related to parts of speech, verb types, and passage length (see McCarthy et al., 2021, for the complete list).

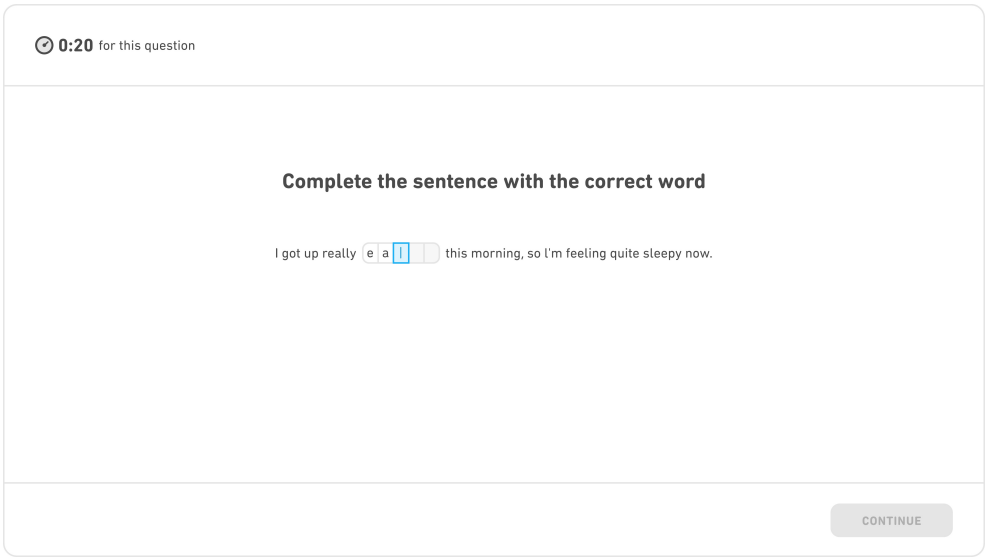

Figure 15. Example Vocabulary in Context Task

Vocabulary in Context

The Vocabulary in Context task type (see Figure 15) measures aspects of vocabulary knowledge relating to meaning, form, and use (Nation, 2001, 2013, 2022). Not only does it assess a wide range of different words (breadth), it also assesses how well a test taker knows different dimensions of these words (depth), and it requires them to access these words in a limited amount of time (fluency; Daller et al., 2007). In terms of the CEFR, this task type relates to the key elements of the Vocabulary range and Vocabulary control scales, and vocabulary knowledge is essential for proficiency in all language skills.

The Vocabulary in Context task type follows the format of the controlled-production vocabulary-levels test (PVLT; Laufer & Nation, 1999), a measure of controlled (as opposed to "free") productive vocabulary knowledge. In our version of this task type, test takers are presented with a sentence that includes a damaged word (the first part of the word is given, and the second part is blank). Candidates for target words are sourced from a large English language reference corpus across four different parts of speech (adjectives, adverbs, nouns, and verbs). The sentences in which the target words are contextualized reflect five different genres (literary, textbook, news, personal writing, and conversation).

Test takers must then complete the word so that it makes sense in the context of the sentence. The damaged portion of the word has a blank space for each character, giving test takers clues about the length of the finished word and constraining possible responses. To ensure that test takers are administered items that tap into multiple aspects of vocabulary knowledge, the test administration requires that all test takers receive at least one item containing either an antonym or a synonym of the target word as a context clue, and at least one additional item containing the target word in a collocation. The time limit for each item is 20 seconds.

Yes/No Vocabulary

"Yes/No" vocabulary tests measure breadth of receptive vocabulary knowledge (Beeckmans et al., 2001). Such knowledge is a critical element of learning a second/additional language (L2) and impacts all language skills, including reading (Laufer, 1992; Roche & Harrington, 2014), listening (Bonk, 2000; Staehr, 2008), speaking (Milton, 2013; Milton et al., 2010), and writing (Kyle & Crossley, 2016; Ruegg et al., 2011). More specifically, yes/no vocabulary tests have been shown to predict reading ability (Milton et al., 2010; Schmitt et al., 2011), as well as writing and listening abilities (Milton et al., 2010). Yes/No vocabulary tests have also been used to assess vocabulary knowledge at various CEFR levels (Milton, 2010).

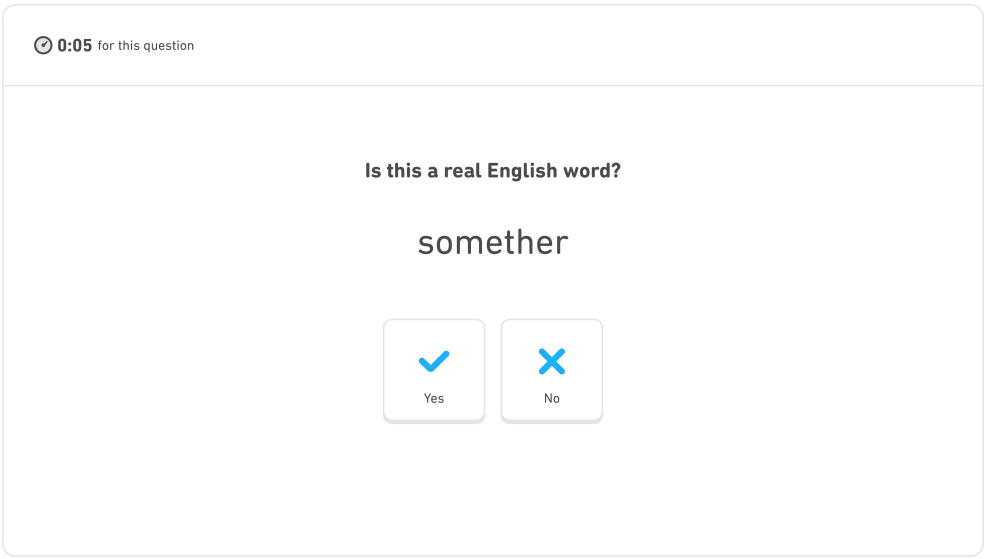

Figure 16. Example Yes/No Vocabulary Task

In the DET version of this task type (see Figure 16), test takers are presented, one at a time, with stimuli that are either a written English word or a pseudo-word designed to appear English-like.* Test takers respond by selecting "Yes" or "No" within five seconds after each word is presented. In order to accurately distinguish real and pseudo-words quickly, test takers need to possess receptive knowledge of a range of lexis, including spelling conventions and morphology. Traditional yes/no vocabulary tests simultaneously present a large set (e.g., 100) of mixed-difficulty stimuli. On the DET, individual yes/no vocabulary stimuli are presented adaptively, based on how the test taker performed on previous items (see Section 5.1 for more on the computer-adaptive administration).

3.4 Listening Tasks

In this subsection we describe the test task types Interactive Listening and Dictation. These tasks primarily involve listening to audio stimuli, although both task types also require test takers to write.

Interactive Listening

The Interactive Listening task type contributes to measurement of the constructs of L2 listening and, through the written summarization sub-task, writing (LaFlair et al., 2023). It complements the Dictation task type (see next section), which focuses more on listening processes, by also measuring aspects of interactional competence. This listening task type is interactive in that it requires test takers to participate in a situationally driven conversation in a university setting. The Interactive Listening task demonstrates correspondence to the TLU domain of English-medium postsecondary studies through the conversation topics, interlocutors (students and professors), and communicative functions, which include asking for clarification about lecture content, making requests, gathering information, asking for advice, planning study sessions, and participating in other university-oriented conversations (Biber & Conrad, 2019).

An Interactive Listening task starts with the Scenario Comprehension sub-task. In this sub-task, test takers hear an aural description of a scenario that describes who the test taker is talking with and for what purpose. After hearing this description, the test taker answers three to four gap-fill comprehension questions about the scenario (See Figure 17. This scenario is followed by

*We use an LSTM recurrent neural network trained on the English dictionary to create realistic pseudo-words, filtering out any real words, including any acceptable spellings, and pseudo-words that orthographically or phonetically resemble real English words too closely.

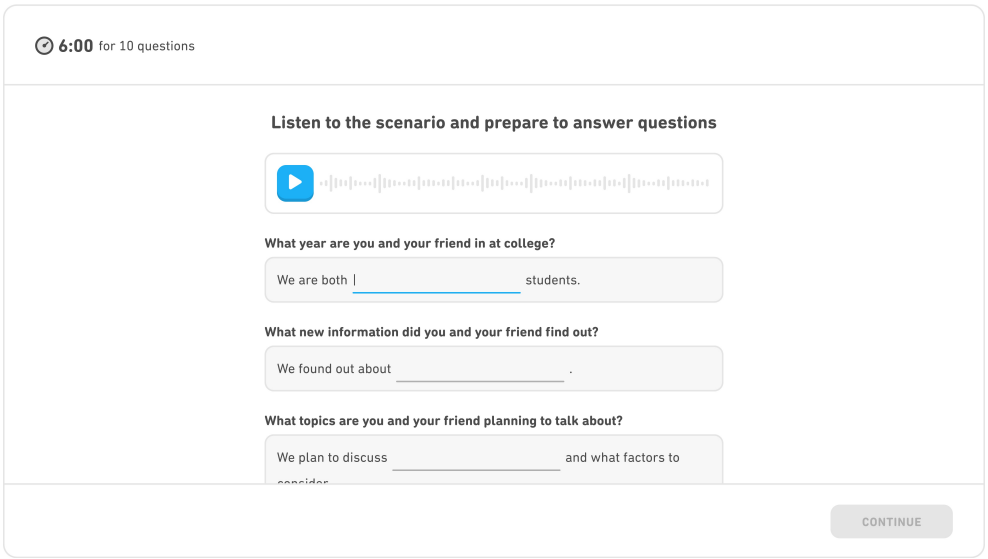

Figure 17. Example Interactive Listening "Scenario Comprehension" Sub-task

the Dialogue Completion sub-task. Some dialogue completions require the test taker to select the first turn in the conversation, while others start with the interlocutor. After each interlocutor turn (which is presented in audio format only), the test taker must select the best response (among multiple options presented in writing) to continue the conversation (see Figure 18). The test taker receives visual feedback after each response; if the response is correct, the box around the text turns green; otherwise, the box turns red, and the correct response is shown. In this way, test takers can respond to the remaining turns based on the intended input. Once the conversation ends, the test taker may use any remaining time to review the conversation before proceeding to the Summarization sub-task (see Figure 19). In the Summarization sub-task, the test taker has 75 seconds to compose a short written summary of the conversation. In contrast to other DET independent writing tasks, the summarization task is an integrated writing task involving accurately and appropriately capturing the main points of the listening in writing.

Each Interactive Listening task exhibits one of three types of conversations: student–student conversations that focus on requests, advice seeking, and other university-oriented purposes; student–professor conversations that focus on similar purposes; and student–professor conversations that focus on information seeking where the student needs to get information about a specific topic from their professor. Each test session includes two Interactive Listening tasks: one student–student conversation and one student–professor conversation. The entire Interactive Listening task lasts 7 minutes and 45 seconds.

Dictation

Dictation is an integrated skills task (listening and writing) that assesses test-taker ability to recognize individual words and to hold them in memory long enough to accurately reproduce them; both abilities are critical for spoken language understanding (Bradlow & Bent, 2002; Buck, 2001; Smith & Kosslyn, 2007). Dictation tasks have also been found to be associated with language-learner intelligibility in speech production (Bradlow & Bent, 2008).

For the DET Dictation task, test takers listen to a spoken sentence or short passage and then transcribe it using the computer keyboard (see Figure 20). Test takers have one minute to listen to the stimulus and transcribe what they heard. They can play the stimulus up to three times. The stimuli used for Dictation items cover a range of language functions including requesting information, expressing opinions, and stating facts. These stimuli also exhibit authentic language features such as contractions and elisions. Given the time constraint and limited number of replays, an understanding of the vocabulary and grammar used in the stimulus is necessary for error-free task completion.

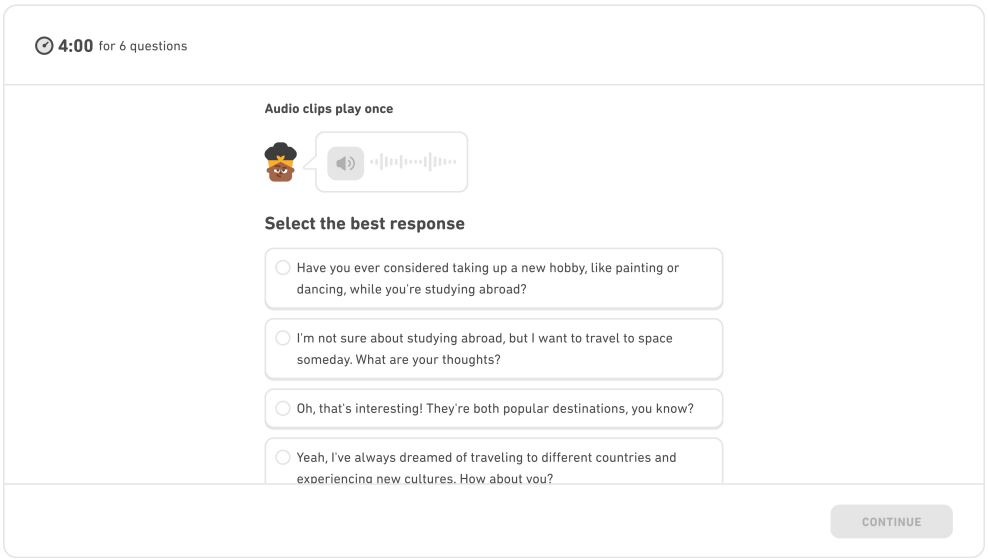

Figure 18. Example Interactive Listening "Dialogue Completion" Sub-task

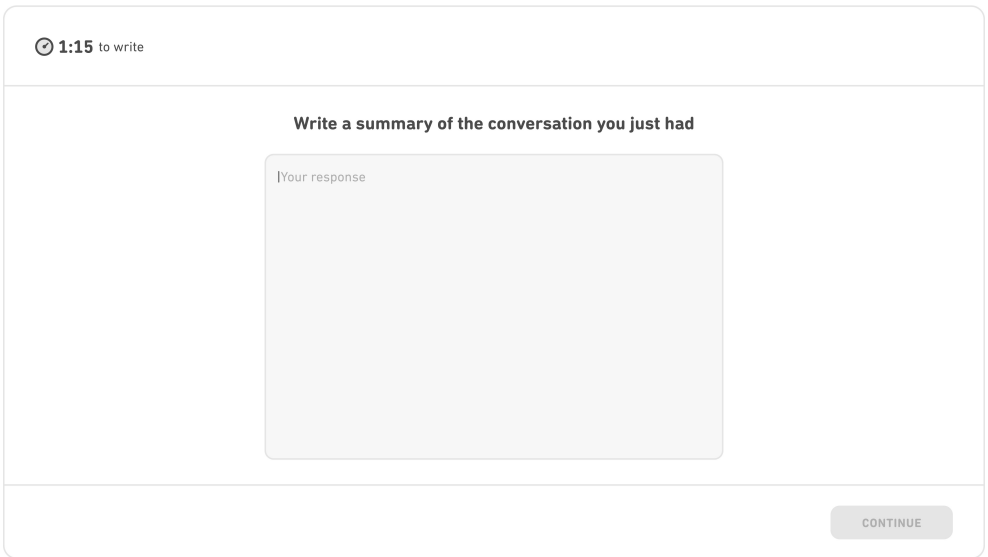

Figure 19. Example Interactive Listening "Summarization" Sub-task

Further readings

- Interactive tests for interactive skills
- The Interactive Reading task: Transformed-based automatic item generation
- Facilitating the writing process on the DET: The Interactive Writing task
- Interactive Listening—The Duolingo English Test

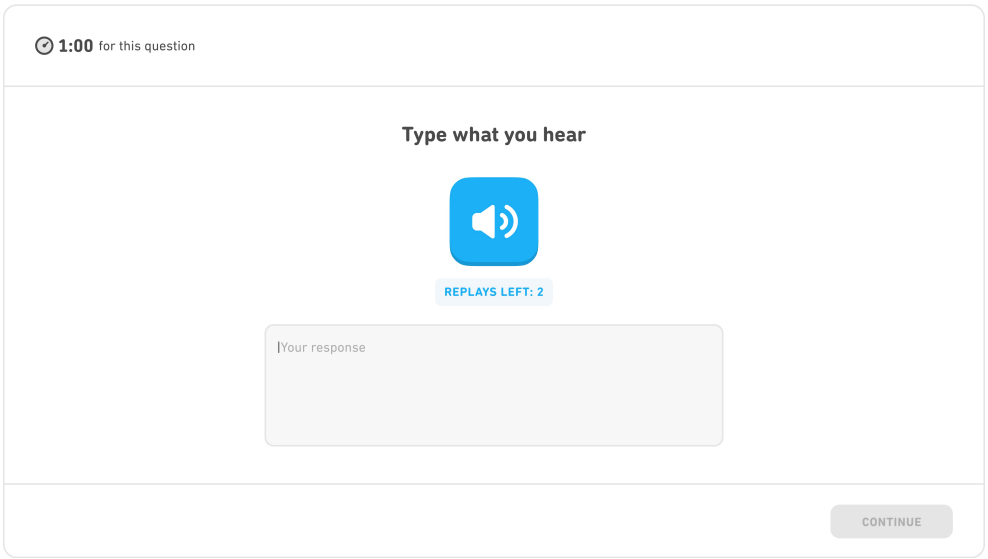

Figure 20. Example Dictation Task

4 Test Development

In this section we describe key test development processes of the Duolingo English Test, including task design, item generation, item review, and item piloting and modelling (test scoring is covered in the next section). See Figure 21 for a simplified heuristic of the test development process. All DET task types are first designed by language assessment experts and refined through iterative field testing. Once task specifications are finalized, item production is scaled up by means of the item factory, which is both a conceptual framework and a system of interconnected software platforms and human review processes used by the DET for scalable high-volume item development (von Davier et al., 2024). This system leverages advanced AI, Natural Language Processing (NLP), and Machine Learning (ML) technologies to automate the initial generation of test items, ensuring a high level of linguistic variety and complexity suitable for a diverse global audience.

The item factory embodies human-in-the-loop AI, in which human oversight and quality control is built into every step (see Burstein et al., 2022). The item factory's human review processes ensure consistency and efficiency while producing items that meet the quality standards for a high-stakes international English language test, including fairness towards all test takers from diverse sociocultural backgrounds (Kunnan, 2024). Involving humans at every stage of item development and review provides additional quality control and oversight, in alignment with the DET's RAI Standards (Burstein, 2025).

4.1 Task Development

All DET task types are designed and approved by language testing specialists, adhering to the principles of the Expanded Evidence-Centered Design (e-ECD; Arieli-Attali et al., 2019) embedded within the DET's theoretical assessment ecosystem (Burstein et al., 2022). Drawing on theoretical background such as the CEFR (Council of Europe, 2001, 2020), assessment researchers create several variants of new task types for field testing to arrive at the final task design. The creation of these task types leverages authentic English-language content, which not only ensures the relevance and real-world applicability of the tasks but also provides input for automatic item generation (AIG). During the task design stage, the task type's AIG process is also developed and refined. Thus, even though all DET task types are designed by experts, AIG allows for the efficient production of a large number of items for each task type, catering to different levels of proficiency and supporting a comprehensive and secure assessment system.

Designs for new task types are tested and refined through a systematic field testing process. New task types or variants are first trialled on the DET's field testing platform, which allows prospective test takers to opt into trying out new task types during the practice test. These new task types are delivered using a randomized experimental design, allowing DET researchers to compare

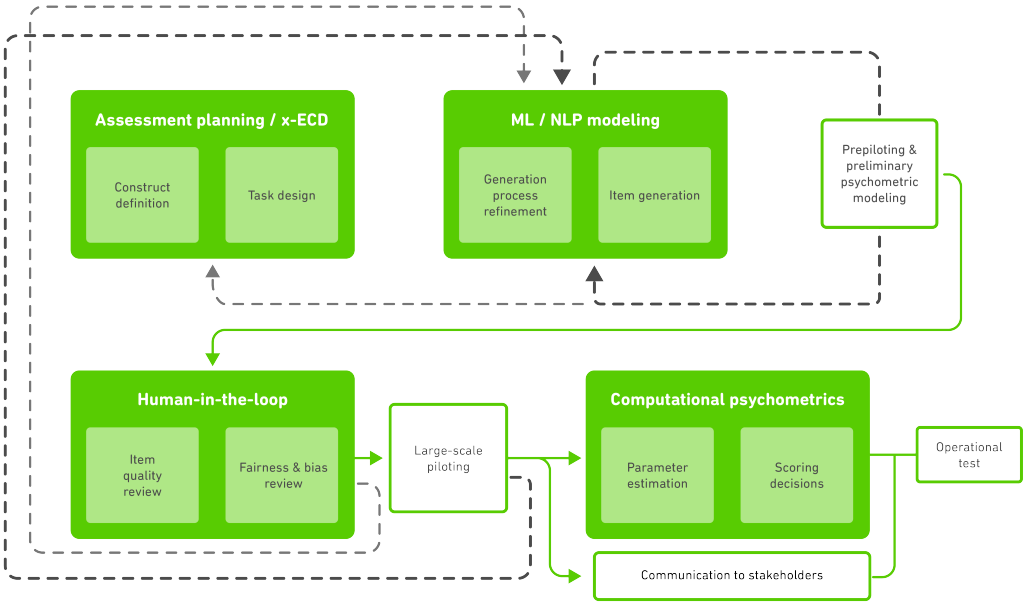

Figure 21. DET Item Generation Process

the performance of new and existing task types, and alternate designs for new task types, under authentic conditions. During this process, specific aspects of the task design are adjusted based on empirical findings – examples include the wording of task instructions, the time limits set for responding, and the number of distractors in multiple choice tasks. Task designs are also evaluated based on psychometric properties, with poorly performing tasks either revised or excluded from operational use. This evidence-driven, human-in-the-loop approach ensures that only task types that meet high standards for measurement accuracy, fairness, and user experience are incorporated into the live DET assessment.

The DET also employs cognitive laboratory studies (aka, "coglabs") to ensure that test takers engage in mental processes analogous to those required in real-world language use. Participants opt in to coglab studies after completing a DET practice test, then complete a fixed set of items while their on-screen actions and verbalizations (i.e., "think-alouds") are recorded. Finally, participants respond to a survey about the processes and strategies they employed to respond to the experimental tasks. Researchers triangulate participants' item responses, post-task surveys, and audiovisual data (screen and microphone recordings) to analyze their approaches to task completion. This multi-method design enables detailed investigation of specific cognitive processes, including orthographic processing (recognizing and spelling words), morphological parsing (analyzing word parts), contextual inferencing (integrating sentence context with lexical knowledge), lexical retrieval, and phraseological processing (e.g., collocation use). Both survey and verbal protocol data reveal whether test takers rely more on automatic recognition or context-driven analysis, offering evidence about the authenticity of cognitive demands. Results are interpreted in light of theoretical frameworks from language testing research. By systematically documenting test takers' response strategies, DET coglabs provide empirical evidence that supports the cognitive validity of test tasks and informs ongoing test development and refinement.

4.2 Item Generation and Review

Each DET item type is generated using unique methods and prompts crafted and iteratively refined by AI engineers. For example, Attali et al. (2022) details the use of Generative Pre-trained Transformer 3 (GPT-3) to generate the passages and response options for the Interactive Reading task. GPT-based AIG involves fine-tuning the prompts to produce items aligned with the task specifications and minimize the proportion of items that are unusable or require manual revision. Once a task's design and generation prompt are finalized, hundreds or even thousands of items can be generated in a short period. Due to the large item

pool made possible by AIG, each test taker only sees a minuscule proportion of existing items, and any two test sessions are unlikely to share a single item. However, not all generated items meet the quality standards for use on the test, so a series of automatic filters and human reviews take place to prevent problematic content from appearing on the test.

Upon generation, items undergo a rigorous review process that combines automated checks with expert human evaluation. This multi-stage review process guarantees that each item meets strict quality standards before being included in any item pool. Automated checks are designed to catch obvious issues of linguistic accuracy and social appropriateness. The automatic filters use language models to detect potentially biased or discriminatory language patterns. Items are also automatically screened for any words or phrases that would not be acceptable in any context, such as terms for specific acts of violence.

Following the automated checks is the item quality review (IQR) stage, in which each item is reviewed (and potentially edited) by subject matter experts (SMEs) to ensure the items are of sufficiently high quality. Written item content is reviewed for adherence to English writing conventions. Audio stimuli are reviewed for accuracy of pronunciation and overall comprehensibility. Item content is also fact-checked to ensure that test takers are not distracted or confused by inaccurate assertions (an example of mitigating the influence of sociocognitive factors on test performance). Both automatic filters and SME reviews are facilitated by a flexible platform and human management system that allow for seamless transitions between phases and collaboration among SMEs.

Next, items go through a sensitivity review referred to as fairness and bias (FAB) review (Church et al., 2025; von Davier et al., 2024). In the FAB review stage, each item is judged by two or three human reviewers against internal FAB guidelines to ensure items are fair towards test takers of diverse identities and backgrounds (e.g., cultural, socioeconomic, and gender). FAB raters are selected to represent diverse identities, perspectives, geographic locations, work contexts, and linguistic backgrounds. As well, all raters have demonstrated experience and interest in promoting equity and diversity. Raters are trained to identify potential sources of construct-irrelevant variance due to either specialized knowledge (e.g., highly technical or culture-specific information) or potentially offensive or distracting content (e.g., cultural stereotypes or descriptions of violence). Items flagged for FAB issues are either edited to resolve the issue or excluded from the item bank if the issue is too extensive. FAB rating data are also used to improve automatic flagging of potentially problematic items. In addition, differential item functioning (DIF) analyses are conducted regularly after the test administrations (Belzak, 2023; Belzak et al., 2023).

The DET's item review platform, called the "item factory", is an internally developed tool for the coordination and oversight of the item review process. The item factory is based on the principles of architecture for smart manufacturing, and it is designed to streamline the creation and review of test items through intelligent automation while incorporating expert human oversight. It assigns items to reviewers and provides reviewers with an interface for editing and rating items. It also tracks editorial suggestions and delivers feedback on the outcomes of such suggestions. Reviewers regularly receive training items interspersed among their assigned items to monitor inter-rater consistency and provide corrective feedback to the reviewers. The item factory facilitates management of the item review process by allowing operations managers to set deadlines and track the progress of items through the process. The item factory also serves a quality assurance function by enabling the monitoring of quality of reviewers' work (e.g., reviewing time and volume, inter-reviewer agreement) and quality of item content (e.g., the number of edits to an item). Access to the item factory is protected with multiple security measures to prevent unauthorized access and maintain the security of the test content.

Further readings

- The item factory: Intelligent automation in support of test development at scale (Chapter 1 in book)
- Guidelines to fair test content: The Duolingo English Test example
- Fairness and justice in language assessment (Chapter 6 in book)

5 Item Delivery and Scoring

This section explains how the computer-adaptive portion of the Duolingo English Test works and how the items are scored. Additionally, it provides information about the automated scoring systems for the speaking and writing tasks and how they were evaluated. For a more detailed overview of the DET's administration and scoring, see Nydick and Lockwood (2024).

Once items are generated, calibrated (i.e., item parameter estimates are made), and placed in the item pool, the DET uses CAT approaches to administer and score tests (Nydick et al., 2024; Segall, 2005; Wainer, 2000). Because computer-adaptive administration gives items to test takers conditional on their estimated ability, CATs have been shown to be shorter (Thissen & Mislevy, 2000), improve TTX for lower-proficiency test takers (Lee & Jia, 2024), and provide uniformly precise scores for most test takers when compared to fixed-form tests (Weiss & Kingsbury, 1984).

The primary advantage of a CAT is that it can estimate test-taker ability (θ) more precisely with fewer test items (Cardwell, Naismith, & Chalhoub-Deville, 2024). The precision of the θ estimate depends on the item sequence: test takers of higher ability are best assessed by items with higher difficulty b_i (and likewise for lower values of θ and b_i). The true value of a test taker's ability is unknown before test administration. As a result, an iterative, adaptive algorithm is required.

5.1 CAT Delivery and Scoring

Each test session begins with a set of Yes/No Vocabulary items followed by Vocabulary in Context items. Table 2 lists the task types administered on each DET session in the approximate order of administration. The task types are loosely organized into the focus areas of "Linguistic resources" (which emphasizes more granular linguistic competences) and "Skills mastery" (which emphasizes more communicative competence). The linguistic-resources task types are administered at the beginning of the test because they take considerably less time per item and therefore provide an estimate of a test taker's proficiency before administering the more time-intensive skills-mastery task types. This measurement efficiency is what allows the DET to be completed in approximately one hour, improving the accessibility of the test.

Table 2. Task types and Administration Order

Task Type	Name for Test Takers	Adaptive	Frequency	Time per Item
Focus area 1: Linguistic resources				
Yes/No Vocabulary	Read and Select	Yes	15–18	0:05
Vocabulary in Context	Fill in the Blanks	Yes	6–9	0:20
C-test*	Read and Complete	Yes	3–6	3:00
Dictation*	Listen and Type	Yes	6–9	1:00
Focus area 2: Skills mastery				
Interactive Reading	Interactive Reading	Yes	2	7-8:00
Interactive Listening	Interactive Listening	Yes	2	7:45
Picture Description (writing)	Write About the Photo	No	3	1:00
Interactive Writing	Interactive Writing	No	1	8:00
Picture Description (speaking)	Speak About the Photo	No	1	1:30
Extended Speaking	Read, Then Speak	No	1	1:30
Interactive Speaking	Interactive Speaking	Yes	1	3:45
Writing Sample	Writing Sample	No	1	5:00
Speaking Sample	Speaking Sample	No	1	3:00

*These item types are interspersed in the computer-adaptive portion of the test

After the initial task, the CAT algorithm makes a provisional estimate of $\hat{\theta}_t$ based on the test taker's responses to time point t. Then the difficulty of the next item is selected as a function of the current estimate: $b_{t+1} = f(\hat{\theta}_t)$. The provisional estimate $\hat{\theta}_t$ is updated after each administered item. Essentially, $\hat{\theta}_t$ is the expected *a posteriori* estimate (Kim & Nicewander, 1993) based on all the administered items up to time point t. This process repeats until a stopping criterion is satisfied.

The CAT approach, combined with concise and predictive task types, helps to minimize test administration time significantly. DET sessions are variable-length, meaning that exam duration and number of items vary across administrations. The iterative, adaptive procedure continues until the test exceeds a maximum length in terms of minutes or items, as long as a minimum number of items has been administered. Almost all test takers complete the DET within an hour (including speaking and writing; excluding onboarding and uploading).

The DET uses a sophisticated scoring process to evaluate test-taker performance. For tasks administered adaptively, which have objectively correct answers, each response is first converted into a numerical grade based on how well it matches the expected answer. Depending on the task, responses may be graded as correct/incorrect or receive a score on a scale that reflects varying levels of accuracy. These grades are then combined with information about the difficulty of each item, using Item Response Theory (IRT) models, to generate an aggregate score (Yancey et al., 2024). Aggregate scores are then used, either alone or in combination with grades from extended speaking or writing items (discussed next), to derive the reported subscores, as explained in Nydick and Lockwood (2024). IRT is a modern approach to test scoring that focuses on the relationship between the test taker's ability and the properties of the test items. It estimates item parameters, typically difficulty and discrimination, based on empirical performance data. By modeling these parameters, IRT provides a way to compare scores even when test takers encounter different sets of items.

5.2 Open-Ended Speaking and Writing Task Scoring

The speaking and writing tasks are scored by automated scoring systems developed by experts at Duolingo in the fields of ML/NLP, computational psychometrics (Burstein & Attali, 2024; von Davier, 2017; von Davier et al., 2021), and applied linguistics: the *Duo Speaking Scorer* and *Duo Writing Scorer*, with separate scoring models for the different task types. These models evaluate each item response based on a number of theoretical speaking and writing subconstructs (i.e., factors contributing to speaking and writing quality).[*] These subconstructs are reflected in human scoring rubrics[†] and are operationalized for automated scoring through the measurement of numerous research-supported linguistic features. Table 3 presents these subconstructs for speaking and writing and provides examples of how these subconstructs are described in both human and automated scoring.

Duo Writing Scorer was evaluated on 2,460 test sessions,[‡] and the *Duo Speaking Scorer* was evaluated on 1,922 test sessions. Each session had 1-2 responses rated by a human rater using the rubrics previously described, and the human ratings for each session were averaged. These raters are experts in the field of English language teaching and assessment and possess the following minimum requirements: five or more years of TESOL experience with adults; TESOL certification (e.g., Cambridge DELTA); relevant bachelor's degree or equivalent; expert English proficiency (C2 on CEFR); and experience in high-stakes speaking/writing assessment. Raters are also selected to ensure a diversity of nationalities, geographic locations, linguistic backgrounds (including English varieties), and work experiences. Raters undergo initial training, must pass a certification test, and are regularly monitored. As part of this monitoring, the DET conducts Many Facet Rasch Measurement (MFRM) analyses to assess the consistency of human ratings and the performance of the rating scales. From these analyses, the score distribution data and category statistics show that the speaking and writing scales are well-functioning and that raters are within acceptable ranges in terms of rater severity for all tasks and skills.

[*] Interactive Speaking tasks are also scored using these same subconstructs and scoring model. However, responses are linked to prompt-specific rubrics containing key content points, and task completion is measured by how many of these points are addressed. Because prompts are selected adaptively based on prior responses, scoring also accounts for variation in prompt content and difficulty.

[†] http://go.duolingo.com/DET_speaking_and_writing_rubrics

[‡] The dataset used for evaluating the *Duo Writing Scorer* pre-dated the introduction of the Interactive Listening summarization task.

Table 3. Open-ended speaking and writing scoring subconstructs

Subconstruct	Example dimensions	Example automated feature
Content	Task achievement, Relevance, Effect on the reader/listener, Appropriacy of style, Development	the cosine similarity between the prompt's embedding and the response's embedding (relevance feature)
Discourse coherence	Clarity, Cohesion, Progression of ideas, Appropriacy of format, Structure (writing only)	predicted coherence rating (0–6 scale) from a language model fine-tuned on SME ratings
Lexis	Lexical diversity, Lexical sophistication, Word choice, Word formation, Error severity, Spelling (writing only)	the proportion of lemmatized words from the response that are level CEFR C1 and above (lexical sophistication feature)
Grammar	Range of structures, Grammatical complexity, Error frequency, Error severity, Appropriacy	the mean tree depth among the dependency trees of each sentence in the response (grammatical complexity feature)
Fluency (speaking only)	Speed, Chunking, Breakdowns, Repairs	number of words per second (speed feature)
Pronunciation (speaking only)	Intelligibility, Comprehensibility, Individual sounds, Word stress, Rhythm, Sentence stress, Intonation, Connected speech	the acoustic model's confidence in the transcription (intelligibility feature)

Table 4 shows the correlation (Pearson *r*) between the average human rating and the automated score (i.e., Human-Machine) and between human ratings for single responses (i.e., Human-Human) for open-ended speaking and writing tasks. Pearson *r* correlations over 0.50 can be interpreted as large (Cohen, 1988). That human–machine and human–human correlations are so similar indicates that the automated scorers are attending to and equally weighting the same subconstructs of speaking and writing as the human raters.

Table 4. Correlation of productive skills tasks by scoring method

	Human–Machine (*r*)	Human–Human (*r*)
Speaking tasks	0.84	0.85
Writing tasks	0.85	0.87

5.3 Calculation of Reported scores

The DET reports four primary subscores—Speaking, Writing, Reading, and Listening—which are derived directly from test-taker responses. These calculated subscores are constructed by combining scores from specific task types associated with each skill (see Table 1 for task-to-score mappings). Depending on the subscore, responses are combined using one of the following approaches:

- IRT modeling: used for Reading, where responses to adaptive tasks are pooled and modeled using expected *a posteriori* (EAP) estimates
- Classical (aggregated) scoring: used for Writing, where grades from open-ended tasks are weighted and averaged directly
- Hybrid scoring: used for Speaking and Listening, where both IRT-based item scores and classically scored tasks contribute to the final subscore

The DET also reports five derived scores, calculated as unweighted averages of relevant subscores:

- Overall Score: the average of the Speaking, Writing, Reading, and Listening subscores
- Literacy: the average of Reading and Writing subscores
- Conversation: the average of Speaking and Listening subscores
- Comprehension: the average of Reading and Listening subscores
- Production: the average of Speaking and Writing subscores

All derived scores are also reported on the 10–160 scale using the same rounding rule. This scoring approach ensures that test takers and test users are able to reproduce composite scores based on simple rules and thus have an intuitive understanding of the meaning of those composites.

For further technical detail on the DET scoring methodology, including specific task weights and model specifications, see the Overview of Duolingo English Test Administration and Scoring report (Nydick & Lockwood, 2024).

Further readings

- An overview of Duolingo English Test administration and scoring
- BERT-IRT: Accelerating item piloting with BERT embeddings and explainable IRT models
- Detecting LLM-assisted cheating on open-ended writing tasks on language proficiency tests

6 Score Comparability and Reliability

Unlike many traditional assessments that rely on simple sum-scoring methods, the Duolingo English Test employs a mix of model-based scoring and task-specific aggregation to estimate test-taker ability (see previous section). This scoring approach ensures that tests are efficient and psychometrically sound, even though each test taker sees a unique set of items using adaptive test methods. As previously noted, responses from individual tasks are pooled and transformed into four primary subscores. These subscores reflect underlying language abilities, rather than raw counts of correct answers.

6.1 Comparability Across Sessions and Versions

The DET maintains score comparability across individual test sessions and test versions through a sophisticated combination of statistical methods, adaptive testing, and equating processes. This approach ensures the interpretability and fairness of reported scores. In each DET session a CAT algorithm selects items from a large, calibrated item pool based on the test taker's estimated proficiency. Consequently, no two test takers see the same set of items, and the difficulty of the items varies across sessions. CAT requires *calibrating* all items using IRT models. As a result, a score of 120 obtained in one session is directly comparable to a score of 120 from any other session, even though the test items differed. This design also enhances test security by minimizing the overlap of content across administrations.

Score comparability across different test versions is ensured through a multi-step process designed to maintain the interpretability of reported scores, even as the test evolves to improve construct coverage and measurement precision. Before any major update, DET researchers analyze extensive data from field testing and piloting of new task types and items. This empirical analysis

informs the scaling parameters needed for the updated test, with the explicit goal of maintaining reported score distributions that align with previous versions. Prior to launching the updated test, DET institutes a code freeze (during which no changes to core test code are deployed) and a marketing freeze (which restricts campaigns that could alter the test-taker population). These freezes minimize differences in the demographic composition of test-taker samples before and after the update, a critical step for accurate comparability analyses. Once the update is live, DET psychometricians compare test-taker scores from sessions immediately before and after the update, examining metrics such as mean scores, variances, and observed score changes of repeat test takers. These post-launch checks verify that any changes introduced by new tasks, items, or scoring logic do not change reported scores for the vast majority of test takers. The intended and typical effect of all DET test updates is to enhance construct coverage and increase score reliability, not to alter the fundamental meaning or interpretation of scores for stakeholders.

The DET sets a validity period of two years for its reported scores. This validity window is grounded in the expectation that an individual's English proficiency may change significantly outside this timeframe. It also provides an upper bound for comparability obligations across test versions: DET's ongoing equating and scaling work ensures that scores earned on any operational version within the past two years remain directly comparable and are interpreted according to the same proficiency standards, regardless of enhancements to the test.

6.2 Differential Item Functioning

We conduct Differential Item Functioning (DIF) analyses to identify test items that are more or less difficult for test takers with different backgrounds (e.g., gender, nationality, age, etc.) despite possessing equivalent levels of English language proficiency. We use general and regularized regression methods to conduct DIF analyses (Belzak, 2023; Belzak et al., 2023). Test-taker background variables are evaluated for DIF independently as well as conditionally on each other. Conditional evaluations of DIF can reduce the likelihood of spurious results.

When an item is flagged for DIF, it undergoes further scrutiny by psychometricians and subject-matter experts to assess the magnitude and implications of the observed differences. Items showing statistically-significant DIF are retained if they are deemed substantively fair and relevant to the construct being measured. Items showing minor DIF effects are also retained if they do not unduly affect the reported scores. However, items exhibiting substantively-large DIF effects with no defensible rationale are either revised or retired from test administration.

6.3 Reliability and Standard Error of Measurement

The reliability of the DET is evaluated by examining multiple scores from repeat test takers (test–retest reliability), as well as the standard error of measurement (SEM). The data used to estimate each of these measures come from a subset of the 467,174 certified tests administered between July 01, 2024 and June 30, 2025.

There are two main challenges with using repeaters to estimate test reliabilities for the full test-taking population. The first is that repeaters are a self-selected, non-random subset of the full testing population. People who choose to repeat tend to represent a more homogeneous, lower-ability subpopulation than the full testing population. Unless addressed, this reduction in ability heterogeneity will tend to artificially reduce estimated reliabilities based on repeaters. The second challenge is that repeaters not only self-select *to* repeat the test, but also self-select *when* to repeat the test. Some repeaters take the test twice in a short period, while other repeaters may wait a year or more to retest. The more time that passes between repeat test takers' sessions, the more opportunity there is for change in test takers' true proficiency. Change in proficiency over time, which varies across individuals, must also be accounted for to avoid artificially reducing reliability estimates.

In order to address the challenges inherent to test–retest reliability, the analysis was conducted on a sample of repeaters who took the DET twice within seven days. The restriction to such repeaters is intended to reduce the impact of heterogeneous proficiency changes on estimated test–retest reliability. To address the fact that repeaters are different from the full population of first-time test takers, DET assessment scientists used Minimum Discriminant Information Adjustment (MDIA; Haberman, 1984). Specifically, MDIA was used to compute weights so that the weighted repeater sample matches all first-time test takers with respect to country, first language, age, gender, computer operating system (Windows vs MacOS), and the means and variances of the DET scores on the first attempt. Weighting in this manner mitigates the potential biasing effects of repeater self-selection on test–retest

reliability estimates (Haberman & Yao, 2015). A weighted test–retest correlation was calculated for the overall score and all subscores. Bootstrapping was used to calculate normal 95% confidence intervals for each reliability estimate. This reliability estimation method is described in greater detail in Belzak (2024).

The point estimates and confidence intervals of the reliabilities for the DET overall score and subscores are shown in Table 5. The subscore reliabilities are slightly lower than the overall score reliability. This finding is expected because subscores are calculated from a smaller number of items. The SEM is estimated based on the standard deviation of the overall score or subscore and the corresponding test–retest reliability estimate. The SEM is a statistic that reflects the accuracy or precision of a score, indicating how much a person's score might vary if the test were taken multiple times. A smaller SEM indicates higher test reliability.

Table 5. Test-Retest Reliability and SEM Estimates (July 01, 2024 — June 30, 2025)

Score	Test–Retest	Lower CI	Upper CI	SEM
Literacy	0.95	0.94	0.95	5.18
Conversation	0.94	0.93	0.94	5.57
Comprehension	0.92	0.92	0.93	6.16
Production	0.95	0.95	0.96	4.87
Speaking	0.94	0.93	0.95	5.68
Writing	0.93	0.93	0.94	6.23
Reading	0.90	0.89	0.91	7.17
Listening	0.87	0.86	0.87	8.73
Overall	0.96	0.96	0.96	4.38

6.4 Added Value of Subscores

In addition to the overall score, the DET reports eight subscores that are also on a scale of 10–160. These eight subscores include four individual skills subscores (Speaking, Writing, Reading, and Listening), which are combined into the four integrated skills subscores: Literacy (Reading and Writing), Conversation (Speaking and Listening), Comprehension (Reading and Listening), and Production (Speaking and Writing). For the task types that contribute to each individual skill (i.e., SWRL) subscore, see Section 3, and for a visualization of the relationship between individual skill and integrated subscores, see Figure 1.

One approach for evaluating subscores is to determine what "added value" they have compared to the overall score (Haberman, 2008). For a subscore to have added value, "the true subscore should be predicted better by a predictor based on the observed subscore than by a predictor based on the total score" (Sinharay et al., 2007, p. 23). In other words, if a reported subscore is less reliable than the overall score, the overall score could actually be a better measure of the subscore's construct.

We thus evaluate each DET subscore for added value using the approach from Haberman (2008), which relies on calculating a proportion reduction in mean squared error (PRMSE). The subscore has added value when its PRMSE is higher than the overall score's PRMSE. In Table 6, we show the PRMSEs for each of the 8 DET subscores with respect to the overall score. Consistent with findings from other high-stakes English language assessments (Sawaki & Sinharay, 2018), Speaking and Production satisfy the added-value criterion. The other subscores do not meet this criterion in part because the DET overall score is extremely reliable.

Notably, the DET reports all subscores to meet stakeholder desires and government requirements, while also informing test score users of these limitations.

6.5 Analytics for Quality Assurance in Assessment

The DET uses a custom-built psychometric quality assurance system, Analytics for Quality Assurance in Assessment (AQuAA; Liao, Attali, von Davier, & Lockwood, 2022; Liao et al., 2021; Liao, Attali, Lockwood, & von Davier, 2022), to continuously monitor test metrics and trends in the test data. AQuAA is an interactive dashboard that blends educational data mining techniques and psychometric theory, allowing the DET's psychometricians and assessment scientists to continuously monitor and evaluate

Table 6. Added Value of Subscores (July 01, 2024 — June 30, 2025)

Score	PRMSE - Overall	PRMSE - Subscore
Literacy	0.98	0.95
Conversation	0.98	0.94
Comprehension	0.98	0.92
Production	0.98	0.95
Speaking	0.90	0.94
Writing	0.97	0.93
Reading	0.95	0.90
Listening	0.98	0.87

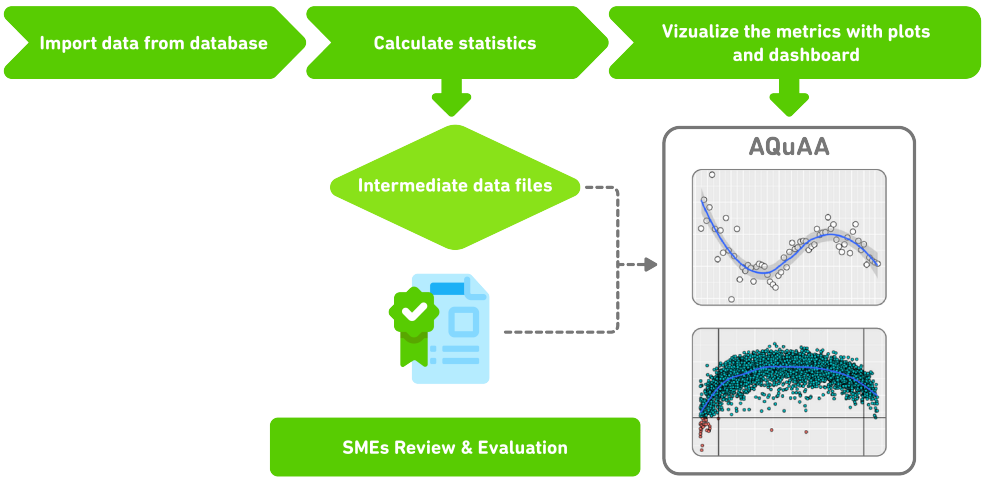

Figure 22. DET Quality Control Procedures

the interaction between the test items, the test administration and scoring algorithms, and the samples of test takers, ensuring scores are consistent over many test administrations. As depicted in Figure 22, test data such as test-taker demographics, item response durations, and item scores are automatically imported into AQuAA from DET databases. These data are then used to calculate various statistics, producing intermediate data files and data visualizations, which are regularly reviewed by a team of psychometricians in order to promptly detect and respond to any anomalous events.

AQuAA monitors metrics over time in the following five categories, adjusting for seasonality effects.

1. **Scores**: Overall scores, subscores, and task type scores are tracked. Score-related statistics include the location and spread of scores, inter-correlations between scores, internal consistency reliability measures and SEM, and correlation with self-reported external measures.
2. **Test-taker profile**: The composition of the test-taker population is tracked over time, as demographic trends partially explain seasonal variability in test scores. Specifically tracked are the percentages of test takers by country, first language (L1), gender, age, intent in taking the test, and other background variables. In addition, many of the score statistics are tracked across major test-taker groups.
3. **Repeaters**: Repeaters are defined as those who take the test more than once within a 30-day window. The prevalence, demographic composition, and test performance of the repeater population are tracked. The performance of the repeater population is tracked with many of the same test score statistics identified above, with additional statistics that are specific to repeaters: testing location and distribution of scores from both the first and second test attempt, as well as their score change, and test–retest reliability (and SEM).

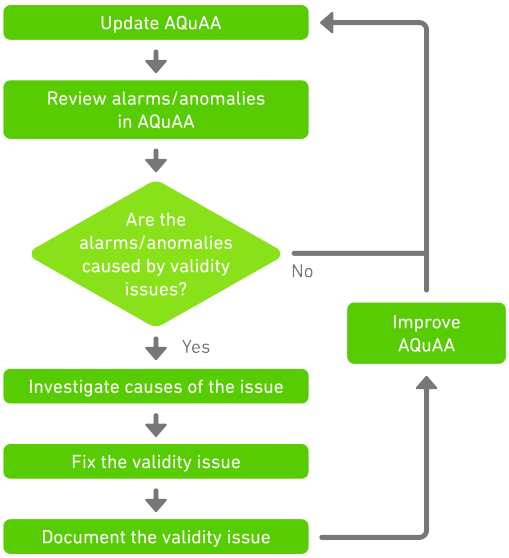

Figure 23. AQuAA Expert Review Process

4. **Item analysis**: Item quality is quantified with four categories of item performance statistics—item difficulty, item discrimination, and item slowness (response time). Tracking these statistics allows for setting expectations about the item bank with respect to item performance, flagging items with extreme and/or inadequate performance, and detecting drift in measures of performance across time.

5. **Item exposure**: An important statistic in this category is the item exposure rate, which is calculated as the number of test administrations containing a certain item divided by the total number of test administrations. Tracking item exposure rates can help flag under- or over-exposure of items. Values of item exposure statistics result from the interaction of various factors, including the size of the item bank and the item selection algorithm.

The quality assurance of the DET is a combination of automatic processes and human review processes. The AQuAA system is used as the starting point for the human review process, and the human review process, in turn, helps AQuAA to evolve into a more powerful tool to detect assessment validity issues. Figure 23 depicts the human review process following every week's update of AQuAA; assessment experts meet to review all metrics for any potential anomalies. Automatic flags have also been implemented to indicate results that warrant closer attention. The assessment experts review any flags individually to determine whether it is a false alarm or further action is required. If the alarm is believed to be caused by a validity issue, follow-up actions are taken to determine the severity and urgency of the issue, fix the issue, and document the issue. Improvements are regularly made to the automatic flagging mechanisms to minimize false positives and false negatives, thereby improving AQuAA's functionality.

While the primary purpose of AQuAA is to facilitate quality control, it also helps DET developers continually improve the exam. Insights drawn from AQuAA are used to direct the maintenance and improvement of other aspects of the assessment, such as item development. Additionally, the AQuAA system itself is designed to be flexible, with the possibility to modify and add metrics in order to adapt as the DET continues to evolve.

Further readings

- Introducing new and updated DET subscores
- AQuAA: Analytics for Quality Assurance in Assessment
- Maintaining and monitoring quality of a continuously administered digital assessment
- AQuAA: Innovative quality control for the future of testing
- Quality assurance in digital-first assessments

7 Relationships With Other Variables

In this section we examine evidence of concurrent and predictive validity for the use of Duolingo English Test scores, for the purposes described in Sections 1 and 2. Specifically, we consider (a) the relationship between DET scores and the scores of other comparable tests of English language proficiency; (b) the relationship between DET content and scores and common language proficiency frameworks; and (c) the relationship between DET scores and subsequent real-world academic performance.

7.1 Relationships With Other Tests

In 2022, a concordance study was conducted (Cardwell, Nydick, et al., 2024) to examine the relationship between DET overall scores and those of TOEFL iBT and IELTS Academic—tests designed to measure similar constructs of English language proficiency and used for the same purpose of postsecondary admissions. The data for this study included the results of certified DET sessions between March 29, 2022 and August 5, 2022, as well as associated TOEFL or IELTS scores for a subset of test takers. A subsequent study was conducted in 2024 to produce concordance tables between the DET and IELTS Academic for the individual subscores (Speaking, Writing, Reading, and Listening). The 2024 study comprised certified DET sessions between April 3, 2024 and September 30, 2024, as well as associated IELTS scores for a subset of test takers.

In both concordance studies, test takers could submit an official score report in exchange for payment or a credit to take the DET again (referred to subsequently as the "official score data"). Prior to any analysis, official score data were assembled, checked, and cleaned by Duolingo assessment scientists and a research assistant. For the 2022 study, in order to achieve recommended minimum sample sizes of 1,500* (Kolen & Brennan, 2004, p. 304) for both TOEFL and IELTS data, as well as to represent a greater range of test-taker ability, the official score data were supplemented with self-report data. (DET test takers have the opportunity to voluntarily report TOEFL or IELTS results at the end of each test session.) For the 2024 study, only official score data were used. Table 7 reports the sizes of the final analytic samples after data cleaning (e.g., removing out-of-range scores and records with invalid subscore–overall score relationships) and restricting the data to DET–TOEFL and/or DET–IELTS score pairs from test dates less than four months apart.

Table 7. Sample Sizes for Correlation and Concordance Analyses

Study year	Scores concorded	Data source	TOEFL	IELTS
2022	Overall	Official	328	1,643
		Self-report	1,095	4,420
2024	SWRL subscores	Official	—	1,943

*This recommended minimum is for the equivalent-groups design. The necessary minimum sample size for a single-group design is theoretically smaller, but a specific number is not given, and so we take 1,500 as the acceptable minimum.

Correlation

Pearson's correlation coefficients were estimated from official score data of the 2022 study to evaluate the relationship between the DET Overall score and those of TOEFL iBT (Table 8). The correlation coefficients show strong, positive relationships between the two tests' scores. These relationships are visualized in Figure 24. The left panel shows the relationship between the DET and TOEFL iBT, and the right panel shows the relationship between the DET and IELTS Academic. Values in parentheses are the sample sizes corresponding to each condition.

Table 8. Correlations Between DET and TOEFL Overall Scores (March 29, 2022 — August 05, 2022)

	TOEFL
All candidates	.71 (328)
Center-based	.82 (183)
Home Edition	.61 (145)

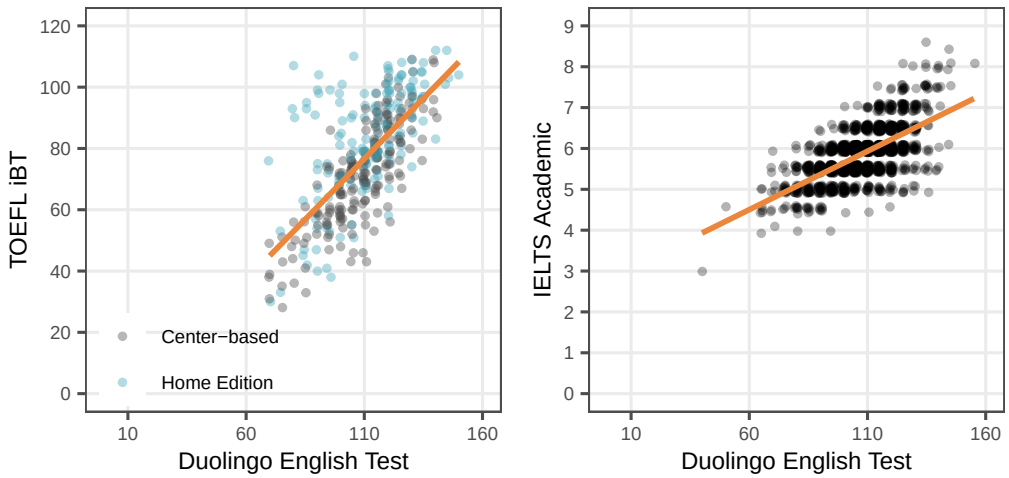

Figure 24. Relationship Between Test Scores

Using official score data from the 2024 study, Pearson correlation coefficients were also estimated between the DET individual skills subscores (Speaking, Writing, Reading, and Listening) and IELTS subscores (Table 9). The observed correlations of both overall scores and subscores support the estimation of concordance tables.

Table 9. Correlations Between DET and IELTS Overall Scores and Subscores (April 03, 2024 — September 30, 2024)

	IELTS
Overall	.73
Speaking	.68
Writing	.54
Reading	.53
Listening	.57

Overall Score Concordance

Given that a sample size of 1,500 is the recommended minimum for building a concordance table using standard equipercentile equating methods (Kolen & Brennan, 2004, p. 304), self-report and official data were both included in the 2022 concordance study to estimate overall score concordance tables. Assessment scientists first used data of individuals who both self-reported a TOEFL

or IELTS score and submitted an official score report to estimate potential reporting bias in self-report data. MDIA (Haberman, 1984) was used to correct for this reporting bias. Follow-up analyses demonstrated that the resulting, adjusted self-report scores had approximately the same statistical properties as the official scores. The DET–IELTS concordance results computed on both the official data and on the combined data were compared to confirm that the combined data set is unbiased. The sample of those who took both the DET and IELTS was sufficiently large to allow for this comparison. After correcting for reporting bias, the self-report and official data were then combined prior to performing the final equating. For individuals with TOEFL or IELTS scores in both the self-report and official score data, only the official score records were retained in the combined data.

Two types of equating were compared in a single-group equating design: equipercentile (Kolen & Brennan, 2004) and kernel equating (von Davier et al., 2004). The equating study was conducted using the equate (Albano, 2016) and kequate (Andersson et al., 2013) packages in R (R Core Team, 2022). Additionally, the data were presmoothed using log-linear models (von Davier et al., 2004) prior to applying the equating methods. The equating methods were evaluated by looking at the final concordance as well as the standard error of equating (SEE), which were estimated via bootstrapping. The final concordance was very similar when comparing equipercentile and kernel equating methods. The standard errors were also very similar across equating methods, although kernel equating had slightly lower and more stable standard errors than equipercentile equating, especially for IELTS given the shorter scale. For these reasons, kernel equating was chosen as the final equating method. See Cardwell, Nydick, et al. (2024) for more details on the methods underlying the DET's concordance tables.

The concordance with IELTS exhibits less error overall because the IELTS score scale contains fewer distinct score points (19 possible band scores between 1 and 9) than the DET (31 possible score values), meaning test takers with the same DET score are very likely to have the same IELTS score. Conversely, the TOEFL scale contains a greater number of distinct score points (121 unique score values), leading to relatively more cases where a particular DET score can correspond to multiple TOEFL scores in the observed data, which inflates the SEE. The overall concordance tables can be found on the DET scores page.

Subscore Concordance

In 2024 a concordance study was conducted to produce concordance tables between DET individual skills subscores (Speaking, Writing, Reading, and Listening) and the corresponding subscores of IELTS Academic. Official IELTS score reports were collected from individuals who had also taken the DET between April 3, 2024 and September 30, 2024 and received certified results. To be eligible for the study, participants must have taken the DET and IELTS no more than 90 days apart. A total of 1,943 usable IELTS score reports were collected, including 591 from participants who had taken DET before IELTS (DET →IELTS) and 1,352 from participants who had taken IELTS before DET (IELTS →DET). The same methods for data cleaning and equating were used as in the previous concordance study (Cardwell, Nydick, et al., 2024). The subscore concordance tables can be found on the DET scores page.

7.2 Relationships With Language Proficiency Frameworks

For many stakeholders, including test takers, teachers, and university admissions officers, it is essential to understand what specific DET scores mean in terms of language use abilities. One way of achieving this goal is to map DET scores onto existing established frameworks of language proficiency, for example the Common European Framework of Reference (CEFR) or the Canadian Language Benchmarks (CLB). Such frameworks help make scores more meaningful and concrete (DeJong & Benigno, 2016) and more generalizable across contexts (Papageorgiou, 2016). In argument-based validation (Chapelle, 2021), this dimension of test score interpretation corresponds to the explanation inference, which considers whether test scores are appropriate indicators of the intended construct (in this case English language proficiency). The DET regularly conducts research to align test scores and language proficiency framework levels, including the CEFR and CLB. Two such recent projects, described below, were presented publicly as part of a symposium at the American Association of Applied Linguistics (AAAL) annual conference (Duolingo, 2025).

The DET–CEFR alignment study (Stethen et al., 2025) was conducted by an independent psychometric company (ACS Ventures, LLC), with consultation from Dr. Anthony Green from the University of Bedfordshire and independent review by the UK's Chartered Institute of Linguists (CIOL). In this study, a diverse panel of twelve subject matter experts reviewed DET test-taker responses and test items across different DET overall score levels, to evaluate whether the DET's threshold levels corresponded to relevant CEFR performance expectations. Overall, the study focused on the extent to which DET tasks reflected relevant CEFR

activities and competencies, how well the DET score ranges aligned with CEFR activities and descriptors, and the consistency of DET scores in relation to CEFR expectations. The results of this study confirmed the appropriacy of the DET's previously established CEFR thresholds. A subsequent similar validation study was conducted in April 2025 to examine the CEFR alignment of DET SWRL subscores, with very similar results to those of the overall score study. For further description of DET–CEFR alignment, see Kostromitina (2024b).

The DET–CLB alignment study (Arias et al., 2025) was organized by the Centre for Canadian Language Benchmarks and carried out by independent experts. The aim of the study was to support DET score interpretation in terms of CLB proficiency levels so that DET scores could be used as evidence of English proficiency in Canadian contexts. The alignment study followed the phases recommended by the Council of Europe (2009) and used the Item-Descriptor Matching method (Ferrara & Lewis, 2012) as it was deemed appropriate for the wide range of tasks and computer-adaptive delivery of the DET. Overall, the alignment process attained satisfactory evaluation evidence to support the interpretation of DET scores in terms of CLB levels. A range of evidence supported the recommended cut scores, including participants' confidence in the standard-setting process, low standard errors of judgment and strong interrater reliability, and triangulation with previous established cut scores from other alignment studies. These findings supported the conclusion that the DET could be used to inform decision-making in the Canadian immigration context.

7.3 Relationships With Real-World Performance

A core goal of language assessment is to ensure that test scores can be meaningfully extrapolated to real-world contexts. In the case of the DET, this means demonstrating that test scores relate to a test taker's ability to succeed in English-medium settings like university classrooms and campuses. In that regard, a growing body of research supports the DET's ability to predict test takers' academic success in higher education settings. A multi-site predictive validity study conducted by the DET research team found no significant differences in GPA outcomes across students admitted to four U.S. universities with DET (n = 541), IELTS (n = 557), or TOEFL (n = 448) scores. The average GPAs for DET students ranged from 3.48–3.72, compared to 3.47–3.75 for IELTS and 3.50–3.78 for TOEFL, demonstrating that DET scores are equally effective at predicting academic success (Kostromitina, 2024a). These results are corroborated by external research. McGehee and Isbell (under review) found that students admitted with DET scores at a large U.S. research university maintained good academic standing and received grades comparable to peers who submitted other English proficiency tests.

The DET's relationship with real-world academic success is further demonstrated in a stakeholder perception study by Isbell et al. (2024). In this study, professors, staff, and students provided ratings of test takers' comprehensibility and academic readiness. These evaluations correlated strongly with DET scores and subscores ($r \geq .74$–$.89$), indicating that higher DET scores are meaningfully associated with perceptions of readiness for academic study, including graduate-level work and teaching assistantships.

Taken together, these findings provide robust support for the validity of the DET, both in terms of predicting academic outcomes and reflecting the language demands of real-world university settings.

Further readings

- Practical considerations when building concordances between English tests
- Is the Duolingo English test aligned with the CEFR?
- DET x CEFR alignment
- How well does the Duolingo English Test predict academic success?

8 Fairness and Impact

Given the Duolingo English Test's mission to lower barriers and increase opportunities for English learners, broad accessibility is one of the central motivations for the test's existence and a primary consideration in any changes to the test. In this section we described how a combination of universally accessible features and accommodations for test takers with disabilities ensures that all test takers have an equal opportunity to demonstrate their English proficiency.

8.1 Access

The DET reflects principles of Universal Design (UD), a framework for designing products and spaces with the goal of maximum accessibility from the start; the concept originated in the field of architecture but has also been adapted to assessment design (Thompson et al., 2002). Maximizing test accessibility through intentional design benefits all test takers, both those with and without disabilities, while simultaneously reducing the need for selective accommodations. The ethos of UD is evident in the origin of the DET and the DET's assessment ecosystem (Burstein et al., 2022), in which all aspects of test design and administration are infused with consideration of the test-taker experience (TTX). The DET's at-home on-demand approach, intuitive user interface, and asynchronous proctoring collectively are designed to reduce physical, socioeconomic, and psychological barriers to test access and optimal test performance.

Perhaps the most salient accessibility benefit of the DET is that at-home testing obviates the need to travel to a physical test center. Traveling to a test center can be burdensome for both socioeconomic and disability-related reasons. Test centers are necessarily concentrated in relatively large urban areas, and some countries do not have any test centers that administer high-stakes ELP tests. It is also not guaranteed that a prospective test taker can obtain a test seat at their closest test center at a time that meets their needs. Many test takers therefore must spend time and money to travel significant distances, even internationally, in order to take a test. This burden is compounded for test takers with disabilities, who might require special transportation or assistance. For such test takers, even local travel can pose a non-trivial barrier. The DET allows most individuals to have their English proficiency evaluated from the most accessible location—their own home.

The AuthaGraph maps (Rudis & Kunimune, 2020) in Figure 25 visualize the issue of physical test access by showing the concentration of test centers in the world (top panel) compared to internet penetration in the world (middle panel), and the concentration of DET test takers (bottom panel; for all tests administered since August 1, 2017). The top two panels of Figure 25 show how much more easily an internet-based test can be accessed than a test center (although Central Africa is admittedly underserved by both models). While the ratio of population to internet access and to test center access is a somewhat limited metric, the potential audience for the DET is clearly orders of magnitude larger than those with access to traditional test centers. As a more specific example, Figure 26 shows the approximate locations of DET test takers relative to test centers currently approved for UK visa purposes. There are test takers on every inhabited continent who would have to travel more than 1,000km to reach the nearest approved test center. By delivering assessments on-demand, 24 hours a day, on any of the world's estimated two billion internet-connected computers, the DET is at the forefront of maximizing test access while maintaining test use validity and test security.

In addition to lowering physical barriers to test access, the DET also embodies accessibility in the economic sense, most obviously through its registration fee, which is a fraction of alternative tests' fees. Furthermore, the DET does not charge extra fees for sharing scores with institutions or appealing proctoring decisions. The DET's at-home on-demand nature removes the need to travel to a test center, potentially representing a cost saving several times greater than the test fee itself. These factors collectively reduce a potentially insurmountable barrier to taking an English language proficiency test, and also make it more feasible for test takers to reattempt the test if needed. The DET's Access Program further reduces socioeconomic barriers for test takers with the greatest need by routinely providing fee waivers to institutions, providing fee waivers to organizations working with populations affected by natural disasters and armed conflicts, and partnering with the UNHCR to provide college counseling to refugee students.

Once test takers have gained access to the DET, the test's design also reduces construct-irrelevant barriers to optimal test performance that could arise during the testing experience. Testing at home gives test takers control over the setup of their testing environment, including the furniture, lighting, and equipment, allowing them to take the test comfortably. This feature is particularly beneficial for test takers with disabilities who may require medical devices or special computer equipment such

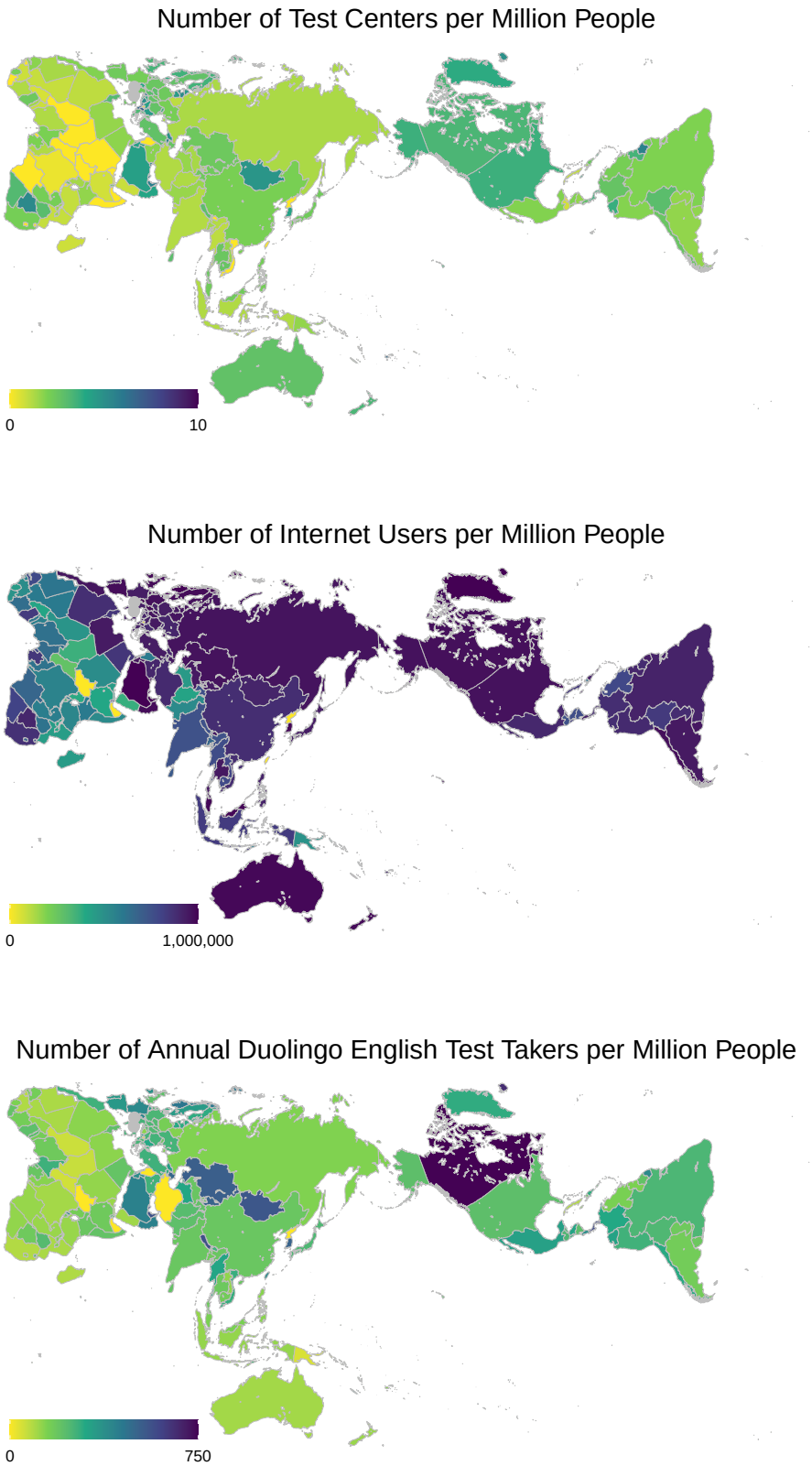

Figure 25. Heatmaps of Test Center Accessibility as of 2025 (top), Internet Accessibility as of 2024 (middle), and Concentration of DET Test Takers in 12 months as of 2025 (bottom)

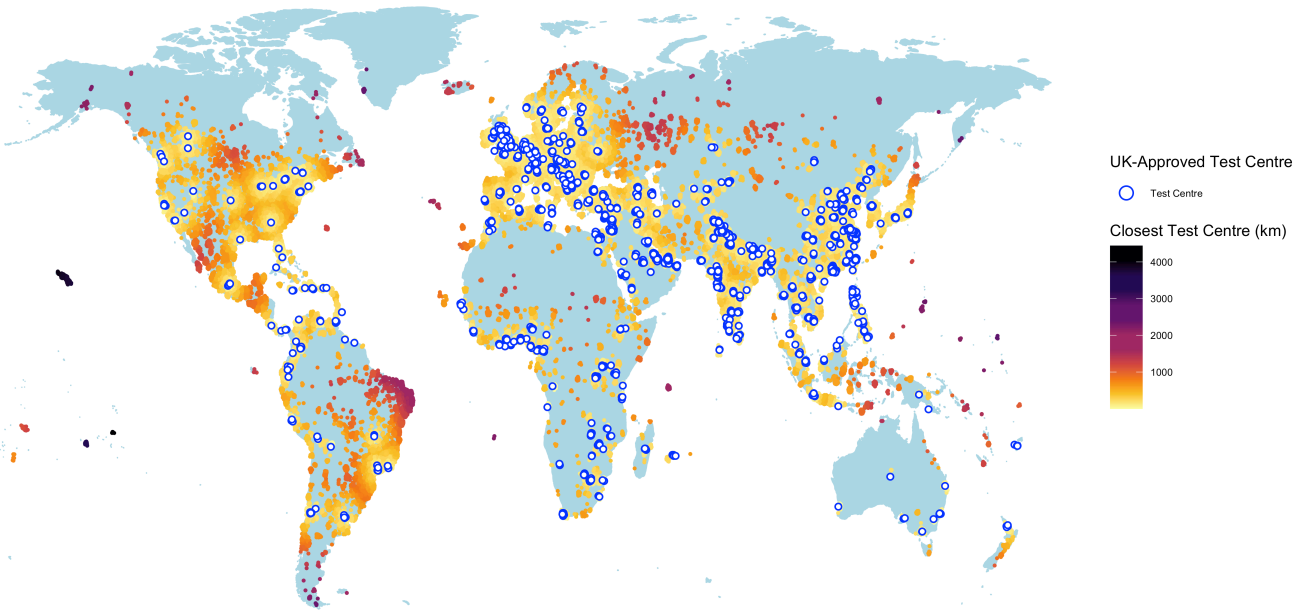

Figure 26. Example: Distance of DET Test Takers from Test Centers Approved for UK Visas

as screen magnification or a special keyboard. The ability to test in a comfortable and familiar environment can also reduce test anxiety (Stowell & Bennett, 2010). The relatively short duration of the test, facilitated by the DET's adaptive nature, may be beneficial for test takers who cannot sit and/or concentrate continuously for long periods due to physical and or psychological disabilities. The DET's user interface complies with W3C Web Content Accessibility Guidelines (WCAG) 2.1 Level AA. Furthermore, the DET's use of asynchronous proctoring (see Section 10.3) likely has a positive impact on TTX, as it does not require interaction with a human proctor and the accompanying concerns about privacy and potential interruptions during testing.

8.2 Accommodations

The DET's inherently accessible design features reduce the need for certain testing accommodations (e.g., extended breaks between test sections). Nevertheless, the DET provides accommodations for both physical (e.g., visual or hearing impairment) and psychological (e.g., autism spectrum disorder) conditions that could constitute construct-irrelevant barriers to optimal test performance. To receive an accommodation, test takers must submit a request at https://englishtest.duolingo.com/accommodations describing both their reason for requesting an accommodation (with supporting documentation, if applicable) and the accommodation requested. The available accommodation options are

- 50% extra time per question
- Accessibility devices (alternate keyboard, etc.)
- Hearing aids
- Headphones
- Listening device
- Screen magnifier/reader
- Other accommodation (to be described by the test taker)

All requests for documented needs are accommodated to the extent reasonable. To ensure accessibility, we have significantly streamlined the process for requesting accommodations compared to the industry standard. The DET requests similar documentation to other English proficiency tests but only requires test takers to fill out a single online form. All inquiries receive a response within three days.

Further readings

- How Duolingo is preparing refugee scholars for university
- Empowering refugee scholars: Duolingo's commitment to accessible education

9 Test-Taker Characteristics

To better understand the context and impact of the Duolingo English Test, this section describes the DET's broad test-taker population that the DET reaches, highlighting the test's focus on fairness and accessibility. The basic test-taker demographics summarized here are based on all certified DET sessions between July 01, 2024 and June 30, 2025. For a more detailed report of test-taker demographics, including distributions of gender identity and age by country[*], see the DET demographic report (Michalowski et al., 2024).

9.1 Demographics

During the onboarding and offboarding process of each test administration, test takers are asked to report their first language (L1), date of birth, reason for taking the test, and their gender identity. The issuing country/region of test takers' identity documents is logged when they show government-issued identification during the onboarding process.

The gender identities of DET test takers are approximately evenly distributed between male and female (Table 10). The median test-taker age is 22. Table 11 shows that 81% of DET test takers are between 16 and 30 years of age at the time of test administration.

Table 10. Percentages of Test-Taker Gender (July 01, 2024 — June 30, 2025)

Gender	Percentage
Female	47.80%
Male	51.04%
Other	0.10%
Not reported	1.05%
Total	100.00%

Test takers are asked to report their L1s during the onboarding process. The most common first languages of DET test takers include Mandarin, Spanish, Arabic, English[†], French, and Portuguese (see Table 12). There are 148 unique L1s represented by test takers of the DET, and the test has been administered to test takers from 219 countries and dependent territories.

Test takers are also asked to optionally indicate their intention for taking the DET, with the choice of applying to a school (secondary, undergraduate, or graduate) and job-related purposes. Table 13 presents the distribution of test-taker intentions.

[*]Note that the *Technical Manual* (this document) and the report *Demographic and Score Properties of Test Takers* use data from different date ranges. Therefore, the reported values may differ.

[†]43% of English-L1 test takers come from India and Canada

Table 11. Percentages of Test-Taker Age (July 01, 2024 — June 30, 2025)

Age	Percentage
< 16	3.24%
[16, 21)	36.09%
[21, 26)	30.45%
[26, 31)	14.40%
[31, 41)	11.42%
≥ 41	4.40%
	0.00%
Total	100.00%

Table 12. Most Frequent Test-Taker L1s (July 01, 2024 — June 30, 2025)

First Language
Chinese - Mandarin
Spanish
English
Arabic
French
Portuguese
Korean
Hindi
Urdu
Nepali

Table 13. Percentages of Test-Taker Intention (July 01, 2024 — June 30, 2025)

Intention	Percentage
Undergrad	47.69%
Grad	33.82%
Secondary School	7.11%
Work	2.93%
None of the Above	8.44%

9.2 Test Performance Statistics

Table 14 provides an overview of the score distributions of the DET from tests administered between July 01, 2024 and June 30, 2025. The descriptive statistics in Table 14 reflect some negative skew, but also indicate that there is no apparent ceiling effect (i.e., the test is not too easy for the target population). The overall score mean and median are 110.59 and 110 respectively, and the interquartile range is 25. For reliability estimates of the overall score and subscores, see Section 6.

9.3 Responsible AI

AI is incorporated throughout the DET, from task development to administration, security, and scoring. The AI applications at each of these stages are guided by the four DET Responsible AI Standards described in (Burstein, 2025). Within each of the standards are descriptions of processes, the goals of which are to mitigate risk introduced by AI and to amplify the opportunities that AI brings to high-stakes language assessments. The first of these standards is validity and reliability. The goals of this standard is to put in place processes that ensure that AI-generated content is appropriate for the purpose of the test (e.g., fairness and bias reviews), that AI scoring is free from bias and has high agreement with human gold standards, and that AI-produced item parameters are evaluated and of high quality. The second standard is fairness. This standard facilitates test-taker access,

Table 14. Descriptive Statistics for Total and Subscores (July 01, 2024 — June 30, 2025)

Score Type	Score	Mean	SD	25th Percentile	Median	75th Percentile
Individual	Speaking	110.96	23.60	95	110	125
	Writing	110.53	23.96	95	110	125
	Reading	109.24	22.87	95	110	125
	Listening	109.15	23.99	95	110	125
Integrated	Literacy	111.13	22.31	100	110	125
	Comprehension	110.45	22.48	95	110	125
	Conversation	111.29	22.24	100	110	125
	Production	111.95	22.49	100	110	125
Overall	Overall	110.59	21.63	100	110	125

accessibility, and inclusion and sets in place processes that ensure demographic representation of test takers in training data for algorithms, such as algorithms used in automated scoring of speaking and writing. Privacy and security is the third standard. This standard outlines processes for data management that ensure test-taker privacy and security and compliance with Duolingo's privacy policy and external policies and laws, such as the European Union's General Data Protection Regulation (GDPR). The fourth standard is accountability and transparency. The processes in this standard focus on documentation of the use of AI in the assessment process and the dissemination of documentation of its use to stakeholders. This comprehensive framework ensures that every application of AI is purposefully managed to uphold the integrity and trustworthiness of the test.

9.4 Test Readiness

The DET provides a comprehensive suite of free test readiness resources designed to familiarize test takers with the test format, instructions, and procedures, thereby minimizing construct-irrelevant variance related to a lack of knowledge about the test. Access to sufficient, high-quality test preparation is an essential element of test fairness. By ensuring that all readiness resources are freely available, including unlimited access to full-length practice tests, the DET supports equity by lowering barriers faced by individuals who might otherwise have limited opportunities for preparation. This commitment is further reflected in the DET's intuitive test design, which aims to reduce the need for rote test-taking strategies or expensive prep courses, factors that can introduce unfair advantages for those with greater resources.

The DET practice test closely simulates the official certified test but draws from a separate item pool and can be taken as many times as desired. Each practice session presents new items, thereby exposing test takers to the range and diversity of question types and formats they may encounter on test day. Validation research has demonstrated that engaging with the practice test increases test takers' confidence and preparation. UX research (e.g., surveys and interviews with prospective and former test takers) informs continual updates and improvements to DET readiness resources. These efforts ensure that DET's approach to test readiness is evidence-based, equitable, and responsive to test-taker needs, contributing to a positive and fair test-taker experience.

Further readings

- Who is taking the DET, and why?
- Duolingo English Test: Demographic and score properties of test takers
- How the DET Practice Test improves test-taker confidence and performance

10 Test Requirements and Security

The Duolingo English Test is administered to test takers via the internet. The security of DET scores is ensured through a robust and secure onboarding process, automated security measures, rules that test takers must adhere to during the test administration, and a strict proctoring process. All test sessions are proctored after the test has been administered and prior to score reporting. Additional security is also provided by the DET's large item bank, CAT format, and active monitoring of item exposure rates, which collectively minimize the probability that test takers can gain any advantage through item pre-knowledge (i.e., exposure to test content before encountering it during an operational test session). Item pre-knowledge is further minimized by preventing repeat test takers (i.e., individuals who take the test more than once) from seeing the same item within a certain period.

Overall, the test security framework is an essential dimension of the larger assessment ecosystem (Burstein et al., 2022), used to protect the integrity of test scores at all stages of the assessment process (LaFlair et al., 2022). The remainder of this section presents a summary of the information found in the Security, Proctoring, and Accommodations whitepaper (Duolingo English Test, 2021) and Duolingo English Test: Security and Score Integrity report (Belzak et al., 2025).

10.1 Test Rules and Procedures

Test takers are required to take the test alone in a quiet environment on a laptop or desktop computer running Windows or macOS and equipped with a front-facing camera, a microphone, and speakers (headphones are not permitted). In addition, a secondary camera (i.e., via a phone or tablet) records their computer and keyboard during the exam. An internet connection with at least 2 Mbps download speed and 1 Mbps upload speed is recommended for test sessions. Test takers are required to take the test through the DET desktop app, which provides a more stable and secure test-taking experience. Test takers are prompted to download and install the desktop app after clicking "Start Test" on the DET website. The desktop app automatically prevents navigation away from the test and blocks tools such as spelling and grammar checkers and automatic word completion.

Before the test is administered, test takers complete an onboarding process. This process checks that the computer's microphone and speaker work. It is also at this time that test takers are asked to show identification and are informed of the test's administration rules, which they must agree to follow before proceeding. In order to ensure test-taker identity, an identity document (ID) must be presented to the webcam during onboarding. An image of the ID is captured.[*] IDs must meet certain criteria, such as being government-issued, currently valid, and including a clear picture of the test taker.

The behaviors that are prohibited during an administration of the DET are listed below. These rules require test takers to remain visible to their cameras at all times and to keep their camera and microphone enabled throughout the test administration. The rules are displayed in the test taker's chosen interface language[†] to ensure comprehension. Test takers are required to acknowledge understanding and agree to these rules before proceeding with the test. If the test session is automatically terminated for reasons such as moving the mouse off-screen or a technical error, a test taker may attempt the test again for free, up to a total of three times. Test takers may contact customer support to obtain additional test attempts in the case of recurring technical errors. Other reasons for test cancellation include:

- Leaving the camera preview
- Looking away from the screen
- Covering ears
- Leaving the web browser
  - Leaving the window with the cursor
  - Exiting full-screen mode
- Speaking when not instructed to do so
- Communicating with another person at any point

[*] ID images are stored temporarily in a highly secure digital repository in compliance with all applicable data privacy regulations and best practices.

[†] Currently available user interface languages: Chinese, English, French, German, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Thai, Turkish, Vietnamese

- Allowing others in the room
- Using any outside reference material
- Using a phone or other device
- Writing or reading notes
- Disabling the microphone or camera

10.2 Automated Security Measures

The DET employs a suite of automated, AI-enhanced security measures that complement human proctoring to uphold the integrity of test scores. These measures analyze multiple data streams, including written responses, user behavior, and audiovisual input, to detect potential violations of test rules and escalate them for human review, all within a human-in-the-loop framework.

The DET incorporates an automated plagiarism detection system that scans open-ended responses for textual overlap with a large corpus of internet content and prior test taker responses. When significant overlap is detected, the system flags the response and provides a visual interface for human proctors to determine whether the similarity is due to common phrasing or substantial duplication. In addition to plagiarism detection, DET's automated security infrastructure detects large language model (LLM) misuse in open-ended writing tasks. This is done through an advanced text analysis system that can recognize when a response was likely written by a generative AI tool like ChatGPT and subsequently copy-typed by test takers. Importantly, the LLM-detector uses contrastive learning and self-training techniques to improve robustness against real-world cheating scenarios, where responses may contain typos or editing from manual transcription (Niu et al., 2024).

In addition to content analysis, the DET records and evaluates behavioral biometrics, such as keystroke dynamics and mouse movement patterns. These signals are used to detect potential proxy test-taking by identifying unusually similar patterns across different test accounts or inconsistencies within the same account over time. This approach is grounded in the principle that such input behaviors are idiosyncratic and difficult to replicate across individuals.

Further automated security measures are provided by automated eye-tracking and sound detection tools. The eye-tracking system visualizes gaze direction across the test session to help proctors identify possible off-screen glances that could indicate misconduct. Similarly, sound detection algorithms analyze test session audio for matches with known notification or device sounds, aiding in the identification of unauthorized device use.

10.3 Human Proctoring

After the test has been completed and uploaded, all DET sessions undergo a thorough proctoring (aka *invigilating*) review by trained human proctors with TESOL/applied linguistics expertise. This review is supplemented by artificial intelligence to call proctors' attention to suspicious behavior. Proctors have access to both audio and video recordings of the entire test session, including both a view of the test taker and a recording of the computer screen. Each test session is reviewed independently by at least two proctors. When necessary, the test session is sent to a third level of review, to be evaluated by a senior proctor or operations manager. This process takes no more than 48 hours after the test has been uploaded. After the process has been completed, score reports are sent electronically to the test taker and any institutions with which they have elected to share their scores. Test takers can share their scores with an unlimited number of institutions. While AI provides assistance at every stage of proctoring, the proctors make the final decision on whether to certify a test. Certain invalid results are eligible to be appealed within 72 hours by submitting a form from the test taker's homepage describing the reason for the appeal. Once the form has been submitted, the test taker will receive an emailed response within four business days informing them of the appeal ruling.

DET proctoring quality is monitored regularly by assessment scientists and subject matter experts (SMEs). A variety of tools and metrics are used to evaluate decision consistency among DET proctors and improve accuracy of decision-making in accordance with proctoring guidelines (Belzak et al., 2024). These tools and metrics include:

Tools

- Monthly reports that track and evaluate proctors' decisions over the last 12 months

  - Used to identify outlier proctors, who then undergo retraining with senior proctors
- Proctor calibration tool that evaluates proctors' decisions using the same test sessions automatically provides immediate feedback about the consensus answer (i.e., what the majority of proctors decide about a test session)
- Calibration meetings between senior and junior proctors, where senior proctors provide feedback on difficult proctoring sessions in a group setting
- Personal training sessions where more experienced proctors shadow less experienced proctors and provide feedback
- Weekly quizzes on proctoring process changes

Metrics

- Percentage of test sessions determined to have rule violations, cheating outcomes, identification issues, or technical errors across time
  - Changes in the test taker population (e.g., due to seasonal trends or market forces) can lead to differences in these trends
- Variability in proctors' decisions across all test sessions proctored, as well as on the same test sessions (e.g., see proctor calibration tool)
- Percentage of decisions overturned between proctors with more and less experience
- Outliers in the percentage of flagged test-taker behaviors, both in terms of under- and overuse (e.g., see monthly reports)
- Average number of minutes taken to proctor a test, controlling for decision type (i.e., rule violation, cheating, etc.) and accuracy of decision
- Test-taker score differences as a function of the type of test-taker behavior that is flagged

The tools and metrics used to monitor proctoring decisions help maintain high-quality, consistent proctoring by continually providing formative feedback to proctors and identifying proctors in need of additional training or re-calibration. Additionally, insights from proctoring quality assurance processes can lead to improvements in test administration and security. For instance, we can identify how and where test takers most often violate rules unintentionally and then modify instructions to minimize rule violation. Maintaining a high degree of consistency across proctors reinforces the security of the DET and ensures that test-taker sessions are reviewed equitably.

Further readings

- Duolingo English Test: Security and score integrity
- Measuring variability in proctor decision-making on high stakes assessments
- Security in action: How proctor shadowing improves the DET
- What is online proctoring?

11 Conclusion

This version of the Technical Manual was produced on July 9, 2025. It provides a detailed overview of all facets of the Duolingo English Test and reports evidence for the DET's validity, reliability, and fairness as outlined in the *Standards for Educational and Psychological Testing* (AERA et al., 2014). Throughout this report, multiple forms of validity evidence are presented to support the use and interpretation of DET scores, focusing on aspects of test content, administration, scoring, and consequences, as well as alignment to language proficiency frameworks and relationships between test scores and test-external variables. The DET is also continually evolving; ongoing research and development efforts are focused on improving all facets of test quality while upholding the DET mission to use assessment technology to lower barriers and increase opportunities for English language learners everywhere. For the latest DET research, visit englishtest.duolingo.com/publications.

12 References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Ahmadi, A., & Sadeghi, E. (2016). Assessing English language learners' oral performance: A comparison of monologue, interview, and group oral test. *Language Assessment Quarterly*, *13*(4), 341–358. https://doi.org/10.1080/15434303.2016.1236797

Albano, A. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software*, *74*(8), 1–36. https://doi.org/10.18637/jss.v074.i08

Andersson, B., Bränberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating with the package kequate. *Journal of Statistical Software*, *55*(6), 1–25. http://www.jstatsoft.org/v55/i06/

Arias, A., Sinclair, J., & Jang, E. (2025). Supporting validity arguments through DET test alignment with Canadian Language Benchmarks [Conference presentation. In A. von Davier, J. Burstein, & A. Verardi (Organizers), *Advancements for multimodal measurement in language assessment: Duolingo & AAAL symposium series*. Presented at the American Association of Applied Linguistics (AAAL), Denver, Colorado.].

Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B., & von Davier, A. A. (2019). The expanded evidence-centered design (e-ECD) for learning and assessment systems: A framework for incorporating learning goals and processes within assessment design. *Frontiers in Psychology*, *10*, 853. https://doi.org/10.3389/fpsyg.2019.00853

Association of Test Publishers. (2024, July). Creating responsible and ethical AI policies for assessment organizations.

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The Interactive Reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.903077

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford University Press.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing*, *28*(1), 51–75. https://doi.org/10.1177/0265532210376379

Barkaoui, K. (2024). Exploring the effects of task difficulty and learner variables on performance on picture description writing tasks. *Assessing Writing*, *60*, 100827. https://doi.org/https://doi.org/10.1016/j.asw.2024.100827

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the yes/no vocabulary test: Some methodological issues in theory and practice. *Language Testing*, *18*(3), 235–274. https://doi.org/10.1177/026553220101800301

Belzak, W. C., Baig, B., Cardwell, R. L., Hastings, R., Horie, A. K., LaFlair, G. T., Liao, M., Niu, C., & Shih, Y.-S. (2025). *Duolingo English Test: Security and score integrity* (Duolingo Research Report No. DRR-25-04) (17 pages). Duolingo English Test. https://duolingo-papers.s3.us-east-1.amazonaws.com/reports/DET_Security_Report.pdf

Belzak, W. C. (2023). The multidimensionality of measurement bias in high-stakes testing: Using machine learning to evaluate complex sources of differential item functioning. *Educational Measurement: Issues and Practice*, *42*(1), 24–33. https://doi.org/10.1111/emip.12486

Belzak, W. C. (2024). Estimating test-retest reliability in the presence of self-selection bias and learning/practice effects. *Applied Psychological Measurement*. https://doi.org/10.1177/01466216241284585

Belzak, W. C., Lockwood, J. R., & Attali, Y. (2024). Measuring variability in proctor decision making on high-stakes assessments: Improving test security in the digital age. *Educational Measurement: Issues and Practice*, *43*(1), 17–29. https://doi.org/10.1111/emip.12591

Belzak, W. C., Naismith, B., & Burstein, J. (2023). Ensuring fairness of human- and AI-generated test items. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky* (pp. 701–707). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36336-8_108

Biber, D., & Conrad, S. (2019). *Register, genre, and style* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108686136

Boers, F. (2018). Picture prompts and some of their uses. *Language Teaching Research*, *22*(4), 375–378. https://doi.org/10.1177/1362168818785219

Bonk, W. J. (2000). Second language lexical knowledge and listening comprehension. *International Journal of Listening*, *14*(1), 14–31. https://doi.org/10.1080/10904018.2000.10499033

Bradlow, A., & Bent, T. (2002). The clear speech effect for non-native listeners. *Journal of the Acoustical Society of America*, *112*, 272–284. https://doi.org/10.1121/1.1487837

Bradlow, A., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*, 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Burstein, J. (2025). *The Duolingo English Test Responsible AI Standards* (Duolingo Research Report No. DRR-25-05). Duolingo. https://go.duolingo.com/ResponsibleAI

Burstein, J., & Attali, Y. (2024). Automated writing evaluation. In A. Kunnan (Ed.), *The Concise Companion to Language Assessment* (pp. 661–670). Wiley.

Burstein, J., LaFlair, G. T., Kunnan, A. J., & von Davier, A. A. (2022). *A theoretical assessment ecosystem for a digital-first assessment—The Duolingo English Test* (Duolingo Research Report No. DRR-22-01). Duolingo. https://go.duolingo.com/ecosystem

Cardwell, R. L., Naismith, B., & Chalhoub-Deville, M. (2024). Computer-adaptive language testing: Focus on language issues. In A. Kunnan (Ed.), *The Concise Companion to Language Assessment* (pp. 649–660). Wiley.

Cardwell, R. L., Nydick, S. W., Lockwood, J. R., & von Davier, A. A. (2024). Practical considerations when building concordances between English tests. *Language Testing*, *41*(1), 192–202. https://doi.org/10.1177/02655322231195027

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge University Press.

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. SAGE Publications.

Church, J., Park, Y., & Burstein, J. (2025). *Guidelines for fair test content: The Duolingo English Test example* (Duolingo Research Report). Duolingo English Test. https://go.duolingo.com/DETFairnessGuidelines

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Coker, D. L. (2012). Descriptive writing. In *Writing* (pp. 159–172). Psychology Press.

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*. https://www.coe.int/lang-cefr

Crossley, S. A., Kyle, K., Allen, L. K., Guo, L., & McNamara, D. S. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation. *Journal of Writing Assessment*, *7*(1).

Daller, M., Milton, J., & Treffers-Daller, J. (2007). Editors' introduction: Conventions, terminology and an overview of the book. In *Modelling and Assessing Vocabulary Knowledge* (pp. 1–32). Cambridge University Press. https://doi.org/10.1017/CBO9780511667268.003

Daller, M., Müller, A., & Wang-Taylor, Y. (2021). The C-test as predictor of the academic success of international students. *International Journal of Bilingual Education and Bilingualism*, *24*(10), 1502–1511. https://doi.org/10.1080/13670050.2020.1747975

de Jong, N., & Vercellotti, M. L. (2016). Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts. *Language Teaching Research*, *20*(3), 387–404. https://doi.org/10.1177/1362168815606161

de Jong, N. H. (2023). Assessing second language speaking proficiency. *Annual Review of Linguistics*, *9*, 541–560. https://doi.org/https://doi.org/10.1146/annurev-linguistics-030521-052114

DeJong, J., & Benigno, V. (2016, November). The CEFR in higher education: Developing descriptors of academic English [Conference presentation. Presented at the Language Testing Forum 2016, University of Reading, UK.].

Duolingo. (2025, March). Advancements for multimodal measurement in language assessment [Symposium presentation. In A. von Davier, J. Burstein, & A. Verardi (Organizers), Duolingo & AAAL Symposium Series. Presented at the American Association of Applied Linguistics (AAAL), Denver, Colorado.].

Duolingo English Test. (2021). *Duolingo English Test: Security, proctoring, and accommodations* (tech. rep.). Duolingo. https://duolingo-papers.s3.amazonaws.com/other/det-security-proctoring-whitepaper.pdf

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, *23*(3), 290–325. https://doi.org/10.1191/0265532206lt330oa

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, *27*(3), 317–334. https://doi.org/10.1177/0265532210363144

Ferrara, S., & Lewis, D. M. (2012). The item-descriptor (ID) matching method. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd Ed.). Routledge.

Goodwin, S., Attali, Y., LaFlair, G. T., Runge, A., Park, Y., von Davier, A. A., & Yancey, K. P. (2023). *Duolingo English Test: Writing construct* (Duolingo Research Report No. DRR-22-03). Duolingo. https://go.duolingo.com/scored-writing

Goodwin, S., & Naismith, B. (2023). *Assessing listening on the Duolingo English Test* (Duolingo Research Report No. DRR-23-02). Duolingo. http://duolingo-testcenter.s3.amazonaws.com/media/resources/listening-whitepaper.pdf

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*(3), 218–238. https://doi.org/https://doi.org/10.1016/j.asw.2013.05.002

Haberman, S. (1984). Adjustment by minimum discriminant information. *The Annals of Statistics*, *12*, 971–988. https://doi.org/10.1214/aos/1176346715

Haberman, S. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204–229. https://doi.org/10.3102/1076998607302636

Haberman, S., & Yao, L. (2015). Repeater analysis for combining information from different assessments. *Journal of Educational Measurement*, *52*, 223–251. https://doi.org/10.1111/jedm.12075

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing (Cambridge Applied Linguistics): Research insights for the classroom* (pp. 69–87). Cambridge University Press.

Isbell, D., Crowther, D., & Nishizawa, H. (2024). Speaking performances, stakeholder perceptions, and test scores: Extrapolating from the Duolingo English Test to the university. *Language Testing*, *2*(41), 233–262. https://doi.org/10.1177/02655322231165984

Iwashita, N., & Vasquez, C. (2015). *An examination of discourse competence at different proficiency levels in IELTS speaking part 2* (tech. rep. No. 2015/5). British Council, Cambridge English Language Assessment, and IDP: IELTS Australia. Melbourne, Australia.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1), 3–17. https://doi.org/10.1177/0265532211417210

Karimi, N. (2011). C-test and vocabulary knowledge. *Language Testing in Asia*, *1*(4), 7. https://doi.org/10.1186/2229-0443-1-4-7

Khodadady, E. (2014). Construct validity of C-tests: A factorial approach. *Journal of Language Teaching and Research*, *5*. https://doi.org/10.4304/jltr.5.6.1353-1362

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika*, *58*, 587–599. https://doi.org/10.1007/BF02294829

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, *14*(1), 47–84. https://doi.org/10.1177/026553229701400104

Koizumi, R., & In'nami, Y. (2024). Predicting functional adequacy from complexity, accuracy, and fluency of second-language picture-prompted speaking. *System*, *120*, 103208. https://doi.org/https://doi.org/10.1016/j.system.2023.103208

Kolen, M., & Brennan, R. (2004). *Test equating methods and practices*. Springer-Verlag.

Kostromitina, M. (2024a). How well does the Duolingo English Test predict academic success? https://blog.englishtest.duolingo.com/does-the-duolingo-english-test-predict-academic-success/

Kostromitina, M. (2024b, October). Is the Duolingo English Test aligned with the CEFR? [Blog post].

Kunnan, A. (2024). Fairness and justice in language assessment. In A. Kunnan (Ed.), *The Concise Companion to Language Assessment* (pp. 80–92). Wiley.

Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12–24. https://doi.org/10.1016/j.jslw.2016.10.003

LaFlair, G. T., Langenfeld, T., Baig, B., Horie, A. K., Attali, Y., & von Davier, A. A. (2022). Digital-first assessments: A security framework. *Journal of Computer Assisted Learning*. https://doi.org/10.1111/jcal.12665

LaFlair, G. T., Runge, A., Attali, Y., Park, Y., Church, J., & Goodwin, S. (2023). *Interactive listening—The Duolingo English Test* (Duolingo Research Report No. DRR-23-01). Duolingo.

Langenfeld, T., Burstein, J., & von Davier, A. A. (2022). Digital-first learning and assessment systems for the 21st century. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.857604

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33–51. https://doi.org/10.1177/026553229901600103

Laufer, B. (1992). Reading in a foreign language: How does L2 lexical knowledge interact with the reader's general academic ability. *Journal of Research in Reading*, *15*(2), 95–103. https://doi.org/10.1111/j.1467-9817.1992.tb00025.x

Lee, Y.-H., & Jia, Y. (2024). Can adaptive testing improve test-taking experience? A case study on educational survey assessment. *Applied Measurement in Education*, *37*(3), 191–208. https://doi.org/10.1080/08957347.2024.2386932

Liao, M., Attali, Y., von Davier, A. A., & Lockwood, J. R. (2022). Quality assurance in digital-first assessments. In *Quantitative psychology: Proceedings of the 2021 meeting of the psychometric society (IMPS)* (pp. 265–276, Vol. 393). Springer. https://doi.org/10.1007/978-3-031-04572-1_20

Liao, M., Attali, Y., & von Davier, A. (2021). AQuAA: Analytics for quality assurance in assessment. *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*, 708–709. https://educationaldatamining.org/EDM2021/virtual/static/pdf/EDM21_paper_79.pdf

Liao, M., Attali, Y., Lockwood, J. R., & von Davier, A. A. (2022). Maintaining and monitoring quality of a continuously administered digital assessment. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.857496

McCarthy, A. D., Yancey, K. P., LaFlair, G. T., Egbert, J., Liao, M., & Settles, B. (2021). Jump-starting item parameters for adaptive language tests. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 883–899. https://doi.org/10.18653/v1/2021.emnlp-main.67

McGehee, M., & Isbell, D. (under review). Relationships between English proficiency test scores and academic outcomes: DET, IELTS, and TOEFL in a U.S. public research-intensive university. https://doi.org/https://osf.io/vguph/download

Messick, S. (1989). Validity. In *Educational measurement, 3rd ed* (pp. 13–103). American Council on Education.

Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1–18). US Department of Education, Office of Educational Research; Improvement.

Michalowski, A., Cardwell, R. L., Nydick, S. W., & Naismith, B. (2024). *Duolingo English Test: Demographic and score properties of test takers* (Duolingo Research Report). Duolingo. https://go.duolingo.com/demographic-score

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 211–232, Vol. 1). EuroSLA.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *Eurosla monographs series 2* (pp. 57–78). European Second Language Association.

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a Foreign Language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83–98, Vol. 52). Multilingual Matters. https://doi.org/10.21832/9781847692900-007

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.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Nation, I. S. P. (2013). *Learning vocabulary in another language (2nd ed.)* Cambridge University Press.

Nation, I. S. P. (2022). *Learning vocabulary in another language (3rd ed.)* Cambridge University Press. https://doi.org/10.1017/9781009093873

Niu, C., Yancey, K., Liu, R., Baig, M., Horie, A., & Sharpnack, J. (2024). Detecting LLM-assisted cheating on open-ended writing tasks on language proficiency tests. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 940–953. https://doi.org/10.18653/v1/2024.emnlp-industry.70

Norris, J. (2018). *Developing C-tests for estimating proficiency in foreign language research*. Peter Lang. https://doi.org/10.3726/b13235

Nydick, S. W., & Lockwood, J. R. (2024). *An overview of Duolingo English Test administration and scoring* (Duolingo Research Report No. DRR-24-03). Duolingo. https://go.duolingo.com/Admin+Scoring

Nydick, S. W., Lockwood, J. R., & Liao, M. (2024). Psychometric considerations for a computerized adaptive language test. In A. Kunnan (Ed.), *The Concise Companion to Language Assessment* (pp. 619–636). Wiley.

O'Sullivan, B. (2008). Notes on assessing speaking. *Cornell University–Language Resource Center*.

Papageorgiou, S. (2016). Aligning language assessments to standards and frameworks. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 327–340). De Gruyter Mouton. https://doi.org/10.1515/9781614513827-022

Park, Y., LaFlair, G. T., Attali, Y., Runge, A., & Goodwin, S. (2022). *Duolingo English Test: Interactive reading* (Duolingo Research Report No. DRR-22-02). Duolingo. https://duolingo-papers.s3.amazonaws.com/other/mpr-whitepaper.pdf

Park, Y., Cardwell, R. L., Goodwin, S., Naismith, B., LaFlair, G., Loh, K., & Yancey, K. (2023). *Assessing speaking on the Duolingo English Test* (Duolingo Research Report No. DRR-23-03). https://duolingo-testcenter.s3.amazonaws.com/media/resources/speaking-whitepaper.pdf

Park, Y., Cardwell, R. L., & Naismith, B. (2024). *Assessing vocabulary on the Duolingo English Test* (Duolingo Research Report No. DRR-24-01). Duolingo. http://go.duolingo.com/vocabulary

Qiu, X. (2022). *International Review of Applied Linguistics in Language Teaching*, *60*(2), 383–409. https://doi.org/doi:10.1515/iral-2017-0094

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Roche, T., & Harrington, M. (2014). Vocabulary knowledge and its relationship with EAP proficiency and academic achievement in an English-medium university in Oman. In R. Al-Mahrooqi & A. Roscoe (Eds.), *Focusing on EFL reading: Theory and practice* (pp. 27–41). Cambridge Scholars Publishing.

Rossiter, M. J., Derwing, T. M., & Jones, V. M. L. O. (2008). Is a picture worth a thousand words? *TESOL Quarterly*, *42*(2), 325–329. http://www.jstor.org/stable/40264459

Rudis, B., & Kunimune, J. (2020). *Imago: Hacky world map geojson based on the imago projection* [R package version 0.1.0]. https://git.rud.is/hrbrmstr/imago

Ruegg, R., Fritz, E., & Holland, J. (2011). Rater sensitivity to qualities of lexis in writing. *TESOL Quarterly*, *45*(1), 63–80. http://www.jstor.org/stable/41307616

Sawaki, Y., & Sinharay, S. (2018). Do the TOEFL iBT® section scores provide value-added information to stakeholders? *Language Testing*, *35*(4), 529–556.

Schleppegrell, M. J. (1998). Grammar as resource: Writing a description. *Research in the Teaching of English*, *32*(2), 182–211. https://doi.org/https://doi.org/10.58680/rte19983904

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*, 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Seedhouse, P., & Harris, A. (2011). *Topic development in the IELTS speaking test* (tech. rep.). IDP: IELTS Australia and British Council. Melbourne, Australia, Manchester, United Kingdom.

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement* (pp. 429–438). Elsevier. https://doi.org/10.1016/B0-12-369398-5/00444-8

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, *26*, 21–28. https://doi.org/10.1111/j.1745-3992.2007.00105.x

Smith, E., & Kosslyn, S. (2007). *Cognitive psychology: Mind and brain*. Pearson/Prentice Hall.

Staehr, L. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*, 139–152. https://doi.org/10.1080/09571730802389975

Stethen, K., Green, A., & Buckendahl, C. (2025). An external validation study for the alignment of an English test to a language framework [Conference presentation. In A. von Davier, J. Burstein, & A. Verardi (Organizers), *Advancements for multimodal measurement in language assessment: Duolingo & AAAL symposium series*. Presented at the American Association of Applied Linguistics (AAAL), Denver, Colorado.].

Stowell, J. R., & Bennett, D. (2010). Effects of online testing on student exam performance and test anxiety. *Journal of Educational Computing Research*, *42*(2), 161–171. https://doi.org/10.2190/EC.42.2.b

Stryker, C., & Kavlakoglu, E. (2024, August). *What is artificial intelligence (AI)?* IBM. https://www.ibm.com/think/topics/artificial-intelligence

Thissen, D., & Mislevy, R. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer (2nd edition)* (pp. 103–135). Routledge.

Thompson, S., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report No. 44). University of Minnesota, National Center on Educational Outcomes. Minneapolis, MN. https://nceo.umn.edu/docs/onlinepubs/synth44.pdf

von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement*, *54*(1), 3–11. https://doi.org/10.1111/jedm.12129

von Davier, A. A., Attali, Y., Runge, A., Church, J., Park, Y., & LaFlair, G. (2024). The item factory: Intelligent automation in support of test development at scale. In H. Jiao & R. W. Lissitz (Eds.), *Machine Learning, Natural Language Processing, and Psychometrics* (pp. 1–26). Information Age Publishing.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. Springer Science & Business Media.

von Davier, A. A., Mislevy, R. J., & Hao, J. (2021). *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in R and Python*. Springer Nature.

Wainer, H. (2000). *Computerized adaptive testing: A primer (2nd edition)*. Routledge.

Wang, G. (2019, October). Humans in the loop: The design of interactive AI systems. https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Weiss, D., & Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*, 361–375. https://doi.org/j.1745-3984.1984.tb01040.x

Yancey, K., Runge, A., LaFlair, G., & Mulcaire, P. (2024). Bert-IRT: Accelerating item piloting with BERT embeddings and explainable IRT models. *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, 428–438. https://doi.org/10.18653/v1/2024.bea-1.35

Young, R. (2011, January). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 426–443, Vol. 2). Routledge.