Interactive Reading—The Duolingo English Test

Duolingo Research Report DRR-22-02
June 16, 2022 (19 pages)
englishtest.duolingo.com/research

Yena Park*, Geoffrey T. LaFlair*, Yigal Attali*, Andrew Runge*, Sarah Goodwin*

Abstract

This paper introduces a new item type on the Duolingo English Test called Interactive Reading and grounds the item within the Duolingo English Test's theoretical language assessment design framework and its assessment ecosystem. The innovative response format and automated item generation methods contribute to the specification of the construct of L2 reading, thereby strengthening the validity claims of the Duolingo English Test.

Keywords

Duolingo English Test, Interactive Reading, reading assessment

Contents

*Duolingo, Inc.

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

1 Introduction

The Duolingo English Test is a digital-first, computer-adaptive, high-stakes proficiency test that assesses English language proficiency for admission to English-medium universities. The Duolingo English Test currently employs twelve different types of items to assess English proficiency in the academic context (Cardwell et al., 2022). Performance on these item types contributes to four subscores (Literacy, Conversation, Comprehension, and Production) and an Overall score. The test is designed to support both efficiency and effectiveness at all stages from development to administration to scoring in large-scale standardized proficiency testing.

Current items that assess reading comprehension on the Duolingo English Test, such as the c-test and read-aloud items, conceptualize the construct of reading comprehension under the psycholinguistics perspective. The emphasis is on test taker-internal cognitive processes that underlie reading, rather than on the product of reading comprehension (Alderson, 2000; Urquhart & Weir, 1998; Van Moere, 2012; Zumbo & Hubley, 2017). Beyond making binary decisions about whether a test taker has correctly understood a text or not, current items measuring reading on the Duolingo English Test elicit the same cognitive processes used in reading (Eskey, 2005; Juffs, 2001; Khalifa & Weir, 2009; Skehan, 1998). A new item that focuses on the product of reading, in addition to the process of reading, enhances the assessment of reading comprehension on the Duolingo English Test.

Interactive Reading is a new item type on the Duolingo English Test that complements the process-oriented perspective of conceptualizing reading with the trait-based perspective (Chapelle, 1999). Interactive Reading presents a passage along with five types of accompanying tasks that include identifying the important ideas and answering comprehension questions specifically geared to assess the level of understanding on the text. These tasks tap into multiple sub-constructs of reading (Grabe, 2009) not only in terms of what they elicit but how they elicit them (Alderson, 2000; Bachman & Palmer, 1996; Qian & Pan, 2014). One example is the response format of highlighting where test takers are asked to answer a comprehension question by highlighting the relevant parts of the text. Altogether the tasks on Interactive Reading help to expand the construct coverage of the Duolingo English Test on academic reading in higher education settings and ultimately strengthen the test validity argument.

The development of Interactive Reading is guided by the assessment ecosystem system (Burstein et al., 2022). The Duolingo English Test ecosystem is a combined network of theoretical frameworks that guide assessment development and evaluation. The ecosystem consists of four different theoretical frameworks: the Language Assessment Design Framework, the Expanded-Evidence-Centered Design (e-ECD) Framework, the Computational Psychometrics Framework, and the Test Security Framework, with considerations for the test-taker experience presiding over the entire ecosystem. The ecosystem contributes to the digitally-informed chain of inferences that supports the test use. The focus of the current paper is the Language Assessment Design Framework that guides how Interactive Reading and its tasks are designed and developed, which will be discussed more in depth in the next section.

This paper introduces Interactive Reading and situates the new item within the Duolingo English Test ecosystem, detailing the construct definition of reading and how the new task

types correspond to and embody the construct definition. The paper ends by revisiting the ecosystem and how it relates to the digitally-formed chain of inferences that supports the test score interpretation and use.

2 Interactive Reading

2.1 Related Work

The construct of L2 reading has conventionally been expressed in several different ways, including conceptualizing reading based on cognitive processes (Alderson, 2000; Khalifa & Weir, 2009), on reading purposes (Britt et al., 2018; Enright et al., 2000; Grabe, 2009), and on the texts in the target-language use (TLU) domain (Green et al., 2010). The Duolingo English Test blends the first two perspectives and envisions the construct of reading both in terms of the purposes with which the test takers read and in terms of the cognitive processes employed while reading (Chapelle, 1999), all in a way that is relevant in academic contexts. All three perspectives are addressed under the Language Assessment Design Framework.

Different response formats have been adopted to tap into the construct of L2 reading. Among many, the discrete-point, selected-response format (for example, the multiple-choice format), has been preferred for the purpose of assessing reading comprehension in high-stakes assessment, research, and classroom settings (Alderson, 2000; Grabe & Jiang, 2014; Qian & Pan, 2014; Riley & Lee, 1996). While recognizing the administrative efficiency of the multiple-choice format and its relevance to the construct (Freedle & Kostin, 1994; Ward et al., 1987), digital-first assessments can actively leverage technology to adopt different response formats that could not be employed in a paper-and-pencil format. Examples of such response formats include highlighting that simulates the act of annotating while reading in the TLU domain. Highlighting is one of the favored reading strategies by university students due to its facilitative role on recall (Kornell & Bjork, 2007; Rice, 1994). It has been shown that highlighting patterns and behaviors are indicative of reading ability and the level of comprehension (Bell & Limber, 2009; Blanchard & Mikkelson, 1987; Winchell et al., 2020) with added benefits for learning (Yue et al., 2015). In other words, what students highlight can reveal what they know and how much they know from the text. Not only are these response formats innovative but they can also help to contribute to a better representation of the construct of reading (Bachman & Palmer, 1996; Qian & Pan, 2014).

2.2 Interactive Reading in the Ecosystem

Each subsection below indicates a component within the Language Assessment Design Framework of the assessment ecosystem and provides detailed descriptions of the theoretical foundation and implementation of Interactive Reading.

2.2.1 Construct Definition

The construct of reading on the Duolingo English Test is defined through reading purposes. Reading purposes not only entail relevant cognitive processes and skills but they are are also most transparent to the stakeholders (Grabe & Stoller, 2020). Table 1 shows different purposes of reading that Interactive Reading taps into as part of the reading construct, how cognitive skills map to each purpose, and brief examples of how each is

instantiated in Interactive Reading (Grabe, 2009; Grabe & Jiang, 2014; Grabe & Stoller, 2020; The Council of Europe, 2020).

Table 1. The construct definition of Interactive Reading

CEFR Categories	Reading Purposes (Grabe, 2009)	Activated Cognitive Skills	Examples
Reading for orientation	To search for information	Search processes Strategic processing abilities	Highlight the Answer
	For quick understanding	fluency and reading speed	
Reading for information and argument	For main ideas	Main-ideas comprehension	Identify the Idea
	To learn	Text-structure awareness Discourse organization	Complete the Passage
	To integrate	Summarization abilities Synthesis skills	
	To use information	Evaluation and critical reading Inferences about text information	Title the Passage

Reading to search for information refers to the purpose of reading to find a specific piece of information within the text. While some consider it as a separate construct from reading comprehension (e.g., Guthrie & Mosenthal, 1987), the ability to search for information has been linked to reading comprehension where better readers are able to better search for information more efficiently and accurately (Cataldo & Oakhill, 2000). Cognitive skills that are involved with the purpose of reading to search for information are search processes and strategic processing abilities.

Reading for quick understanding involves skimming a text (or parts of a text) to form a general understanding of the text from limited information. It involves cognitive skills such as automaticity, fluency, and reading speed (Guthrie, 1988; Guthrie & Kirsch, 1987). Reading to search for information and reading for quick understanding are covered under Common European Framework of Reference (CEFR) as reading for orientation.

Reading to search for information and *reading for quick understanding* are invoked most frequently for university students who rely heavily on the Internet to search for and select sources that align with specific goals (Grabe & Stoller, 2020; Head & Eisenberg, 2009; Thompson et al., 2013).

Reading for main ideas involves general reading comprehension of main ideas that underpins all reading purposes and activities. An example of cognitive skills involved in reading to understand is main-ideas comprehension.

Reading to learn requires readers to understand how ideas within a text connect to each other and to readers' prior knowledge. Readers not only comprehend the main ideas and details of a text, but also store them in a coherent, organized fashion for uses that extend beyond general comprehension. Cognitive skills involved here are text-structure awareness and discourse organization.

Reading to integrate requires readers to combine information from multiple texts, or different parts of a long text. This requires building a larger frame of organization under which the discourse structure of each text (or each part of a text) must be filed. Cognitive skills for reading to integrate information involve summarization abilities and synthesis skills.

Reading to use information refers to the purpose of reading to extract relevant information from a text (or multiple texts) and apply the carefully curated information, combined with background knowledge, to interpret the text or perform other tasks. Cognitive processes involved during reading to use information are inferences about text information, evaluation, and critical reading.

Reading for main ideas, *reading to learn*, *reading to integrate*, and *reading to use information* are addressed in CEFR guidelines as reading for information and argument. *Reading to learn*, *reading to integrate*, and *reading to use information* in addition are often carried out in academic settings (Grabe, 2009).

In addition to the domain-specific construct of reading outlined above, the Duolingo English Test ecosystem framework utilizes the sociocognitive framework to subsume ancillary skills that are relevant to successful language use in academic context under the construct of language proficiency. These sociocognitive factors include secondary constructs such as integrated skills, pragmatic skills, and interactional skills; general skills such as critical thinking and content knowledge; and other influential factors such as intrapersonal factors, experiential factors, and neurological factors.

2.2.2 Task Types

This subsection describes how each task type within Interactive Reading is designed to elicit parts of the construct outlined in the previous subsection. Most tasks cover more than one purpose of reading at the same time but only the primary purpose that is relevant to each task type is discussed.

Interactive Reading consists of five tasks:

1. Complete the sentences
2. Complete the passage
3. Highlight the answer
4. Identify the idea
5. Title the passage

Performance on these tasks contributes to Literacy and Comprehension subscores, as well as the Overall score.

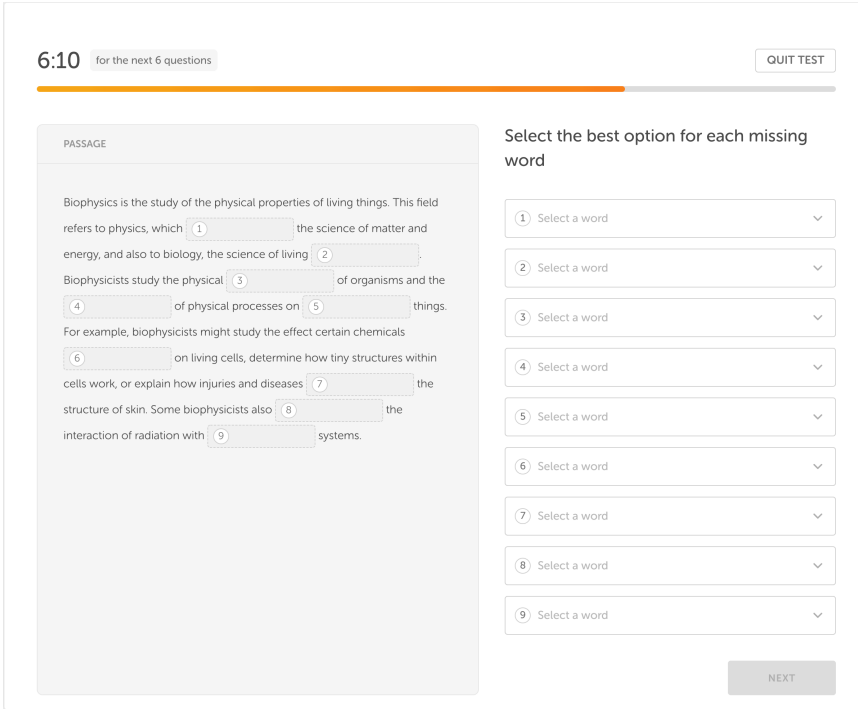

Figure 1. In the Complete the Sentences task, test takers see only a part of the reading passage with deleted words.

2.2.2.1 Complete the Sentences In the *Complete the Sentences* task, only the first half of the reading passage is displayed (see Figure 1 and Figure 2). Test takers are asked to choose the most appropriate word for the blank from five options.

The *Complete the Sentences* task fulfills the purpose of reading for quick understanding and reading for general comprehension, primarily engaging lexical, morphological, and syntactic knowledge. Reading comprehension actively interacts with linguistic skills during this task where linguistic knowledge is required to support the level of comprehension needed to understand the passage which in turn is used to find the most appropriate word for the blanks in the passage (Alderson, 2000). Lexical knowledge is critical for fluent reading (Grabe, 2009) with implications for reading competence (Cheng & Matthews, 2018; Milton, 2013; Qian, 2002; Qian & Schedl, 2004).

2.2.2.2 Complete the Passage The *Complete the Passage* task reveals the full passage with one sentence missing (see Figure 3 and Figure 4). Test takers are asked to choose, from a series of options, the best sentence that completes the passage.

The *Complete the Passage* task represents, to a limited extent, the purpose of reading to learn and to integrate information. While the purposes of reading to learn and to integrate information

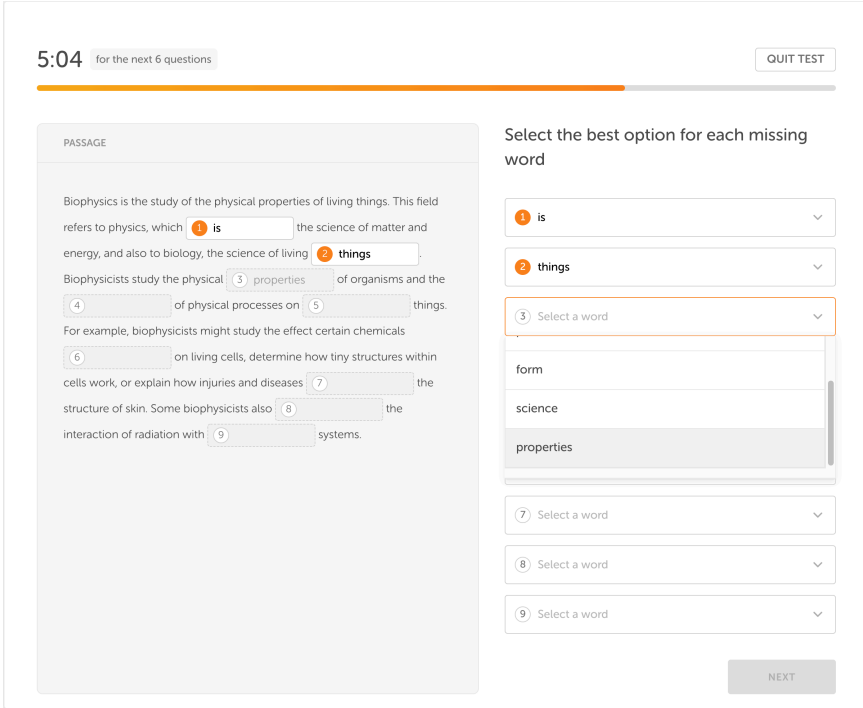

Figure 2. In the Complete the Sentences task, test takers are asked to choose from a list of options the most appropriate word for each blank.

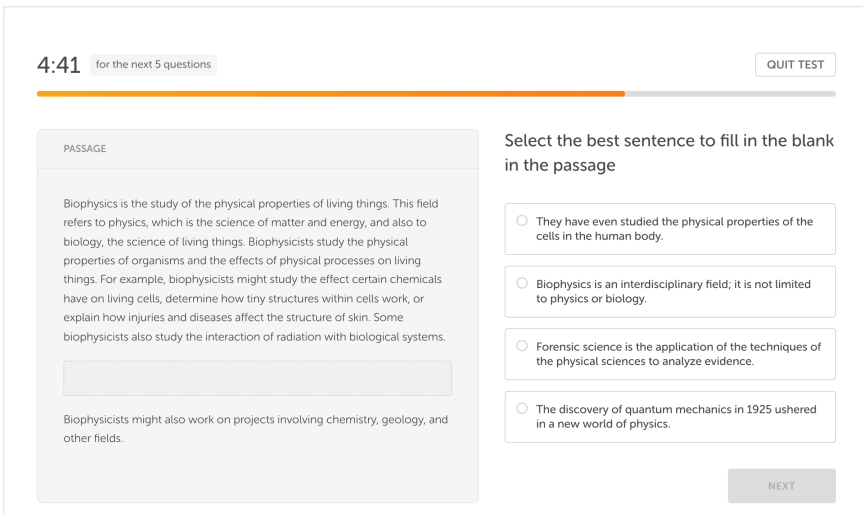

Figure 3. In the Complete the Passage task, test takers see the full passage, with a deleted sentence.

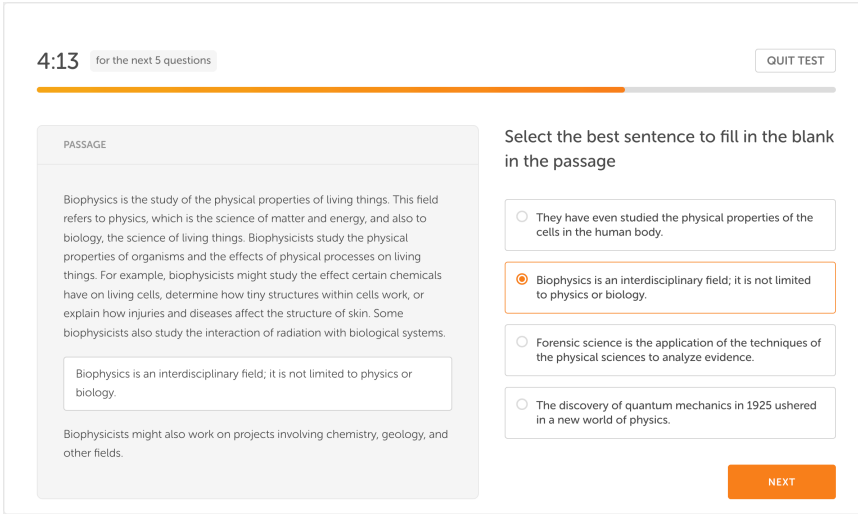

Figure 4. In the Complete the Passage task, test takers are asked to find a sentence that best completes the passage.

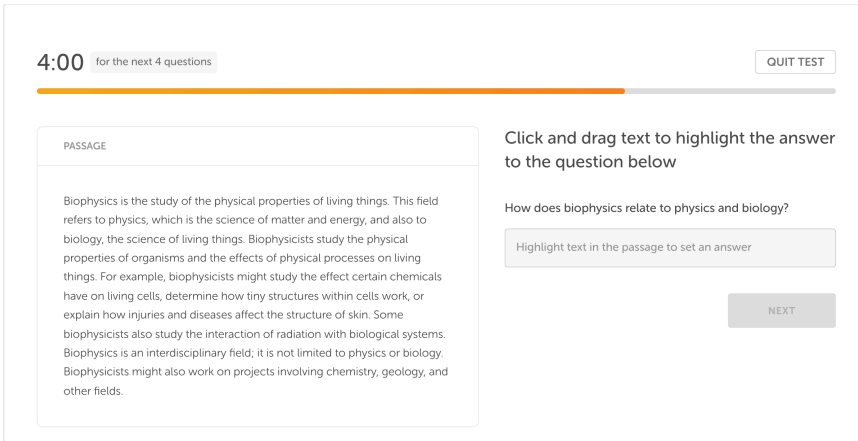

Figure 5. In the Highlight the Answer task, test takers are asked two comprehension questions.

are assumed to be applicable mostly to longer texts, these can be applied in a limited fashion to the *Complete the Passage* task in that it requires test takers to infer how the two halves of a passage connect to one another. In doing so, test takers must recognize the discourse cues of each half and reconcile the two to arrive at the larger rhetorical structure, which they assemble themselves with the best sentence connector from the given options.

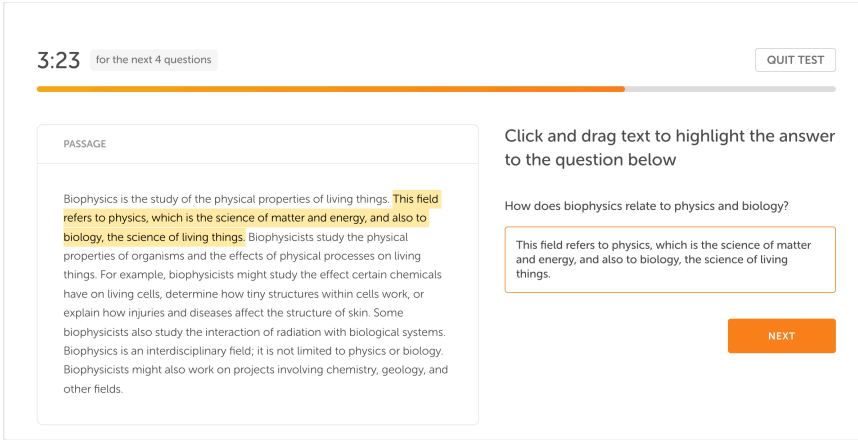

Figure 6. In the Highlight the Answer task, test takers are asked to highlight the parts of the text that answer the comprehension questions.

2.2.2.3 Highlight the Answer

The *Highlight the Answer* task reveals the full passage with two reading comprehension questions (see Figure 5 and Figure 6). Test takers are asked to highlight the parts of the text that would answer the comprehension questions.

The *Highlight the Answer* task represents the purposes of reading to search for information and reading for quick understanding. The task activates major component abilities of reading such as search processes depending on the types of questions asked. The purpose of reading to search for information is instantiated and targeted by the response format that asks test takers to highlight the relevant parts in the text, where the response format simulates the act of executing on the purpose of reading to search for information. This particular response format addresses the issue of underrepresentation of reading to search for information in large-scale standardized proficiency tests (Grabe & Jiang, 2014). In addition, the response format of highlighting is highly faithful to the reading behavior in the TLU domain (Rice, 1994), as well as being indicative of reading ability, comprehension and efficiency (Bell & Limber, 2009; Blanchard & Mikkelson, 1987; Winchell et al., 2020).

2.2.2.4 Identify the Idea

The *Identify the Idea* task asks test takers to choose an idea that is expressed in the passage (see Figure 7). This can either be a detail in the passage, or a main idea of the passage.

The *Identify the Idea* task addresses the purpose of reading to understand. The task engages cognitive skills such as comprehension of important details and main-ideas.

2.2.2.5 Title the Passage

The *Title the Passage* task asks test takers to choose the best title for the passage (see Figure 8).

The *Title the Passage* task fulfills the purpose of reading to use information. The task asks test takers to recall their prior notion of what constitutes a good title and use this in conjunction with what they gleaned from the passage to identify the best title that captures the essence of the

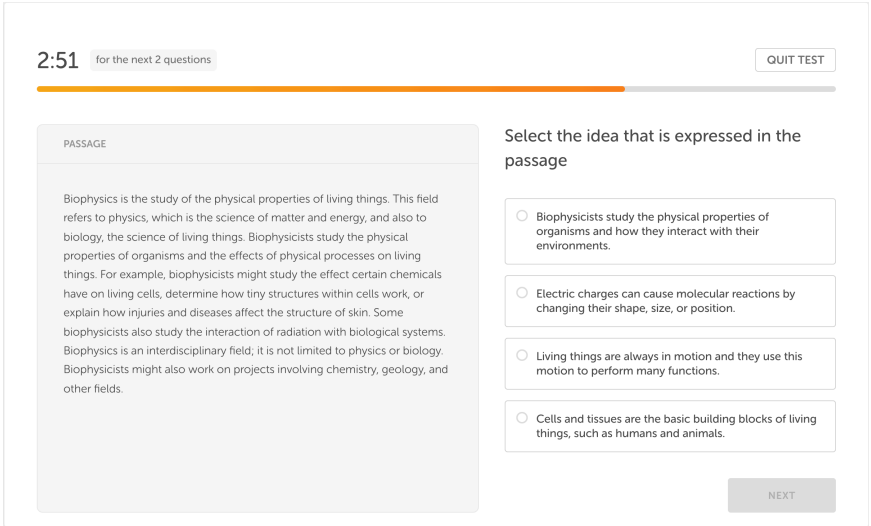

Figure 7. In the Identify the Idea task, test takers are asked to select an idea that is expressed in the passage.

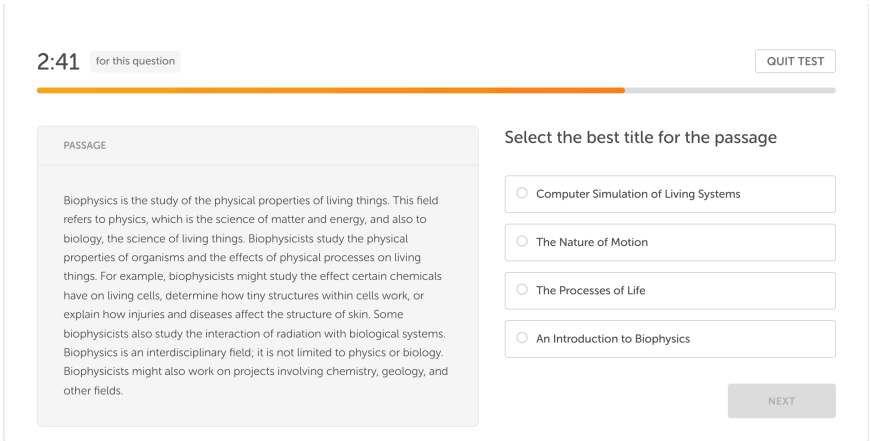

Figure 8. The Title the Passage task, test takers are asked to choose the best title for the passage.

passage. This task activates skills such as evaluation and critical reading, inferences about text information, and summarization abilities.

Interactive Reading also considers socio-cognitive factors that may interfere with performance and actively addresses these to mitigate the possibility of interference with accurate score interpretations (Burstein et al., 2022). For instance, texts that focus heavily on one specific subject have the potential to favor test takers who are more knowledgeable about that particular

subject (Brantmeier, 2005; Clapham, 1998; Krekeler, 2006). This is addressed through an extensive review of the items for fairness and bias issues by a human panel of reviewers with backgrounds in language teaching and linguistics. Intrapersonal and experiential factors that affect test takers are mitigated through readily available test-readiness resources like free, unlimited practice tests. Neurological factors are addressed through user experience testing and multiple pilots to determine the time allotted. Additionally, user experience testing prior to the implementation ensures that tasks are designed and delivered in a way that is accessible to all test takers.

2.2.3 Automated Item Generation and Scoring

2.2.3.1 Passages

All reading passages and accompanying items (including the stems and the distractors for tasks using the multiple-choice format) are automatically generated by Generative Pre-trained Transformer 3 (GPT-3). GPT-3 excels at few-shot learning, which means it can be given a small number of representative samples (i.e., narrative and expository passages) of text in order to complete a task, such as text generation (Brown et al., 2020). Passages in Interactive Reading are generated to reflect the types of texts that university students typically encounter in the TLU domain, lending support to using the Duolingo English Test for higher education admission purposes.

The texts that are automatically generated for Interactive Reading feature two major categories of texts: expository and narrative, which are representative of the TLU domain. Open access texts from registers, such as textbooks and news articles have been used as prompts to generate novel texts that are representative of expository language in academic and non-academic domains. Textbooks are a popular source of information for university students (Thompson et al., 2013; Weir et al., 2009), whereas news articles are important for university students in everyday life (Head & Eisenberg, 2009). Similarly, narrative prompts are supplied to GPT-3 as reference texts for generating large batches of novel narrative reading passages. Narrative recounts are commonly used in academic texts, such as ethnographic reports, reflection, and biography (de Chazal, 2014). All these text types represent the texts typically encountered in the TLU domain.

The passages in Interactive Reading undergo three stages of quality review after automatic generation. The first stage is an automated screening stage where passages that do not meet the predetermined criteria are excluded. Some of the criteria are:

- Minimum/Maximum number of sentences
- Minimum/Maximum number of words
- Minimum/Maximum number of characters
- Duplicated words/phrases/sentences
- Presence of extremely rare words
- Presence of potentially offensive/inappropriate words/phrases/sentences
- Punctuation or grammatical errors
- Difficulty estimated by an external machine learning model
- Estimates of the approximate average likelihood of any phrase or sentence in the passage

The second stage involves minor editing by human reviewers to improve the flow of the passages. The third stage consists of a human review of fairness and bias issues. Each passage is read

by human reviewers to evaluate the subject matter and content for fairness and potential bias. This specifically includes screening passages, items, and options for any controversial and problematic topics as well as topics that may not be accessible to international test takers. All reviews work to ensure both the delightful test taker experience and the assumption that what the Duolingo English Test measures is free of interference from what it does not intend to measure.

2.2.3.2 Items Automatic item generation for Interactive Reading involves generating options (both the correct answers and distractors) for tasks with the multiple-choice format and generating questions for the reading comprehension task. Table 2 describes how each task type in Interactive Reading is automatically generated.

Automatic item generation allows the generation of multiple correct options and distractors, which are then evaluated and selected first automatically based on a set of criteria and then by a panel of human reviewers with item development experience. Samples of such criteria are shown in Table 3.

2.2.3.3 Grading Interactive Reading uses two methods to grade the responses: binary and partial credit. Complete the Sentences, Complete the Passage, Identify the Idea, and Title the Passage adopt the multiple-choice format and consequently binary grading for each item. The Highlight the Answer task is graded based on the distance between the text highlighted by a test taker and the correct response. This is calculated as the Euclidean distance between the start- and end-points of the provided and expected responses. These scoring methods allow all tasks in Interactive Reading to be scored automatically, supporting the adaptive nature of the Duolingo English Test and its concomitant large-scale test development and administration.

2.2.4 Evidence Specification Interactive Reading collects binary and continuous response data to build a score profile for how much a test taker has understood from the passage. Interactive Reading is currently not using process data; more research is needed on the relationship between process data (such as response time) and proficiency to warrant its inclusion (Zumbo & Hubley, 2017).

Preliminary data for the evidence specification stage comes from a series of pilots that was administered at the end of the practice test (see 2.2.5). Scores on Interactive Reading reported moderate correlations with c-test and read-aloud items; they also showed moderate correlations with self-reported subscores of reading on other large-scale high-stakes standardized English proficiency tests.

A large-scale pilot was conducted for 21 days with 454 passages and a total of 5,246 items. A total of 425 responses were collected per item. The items were overall widely distributed in their easiness with an overall facility value of 0.70. Item-total correlations demonstrate the discriminatory power of Interactive Reading. The items in Interactive Reading showed reasonably moderate to high discrimination, with an overall average of 0.27. Analyses were performed to remove distractors with lower discrimination indices to improve the overall discriminatory power of items.

The results of the pilots of Interactive Reading have demonstrated that these items have met the minimum requirement for subsequent, more complex psychometric modeling where they will

Table 2. The automated item generation methods of Interactive Reading

Task	Item Generation	
	Correct Option	Distractor
Complete the Sentences	Words to hide based on: • The most likely word by BERT (Bidirectional Encoder Representations from Transformers; Devlin, 2018) • Linguistic analyses (lexical, syntactic, and context)	• BERT likelihood • Lexical and syntactic analysis
Complete the Passage	A sentence to elide is chosen based on: • The sentence's likelihood from the preceding sentences • The average likelihood of the following sentences • The number of words in the sentence	Sentences chosen from alternative passages
Identify the idea	GPT-3 uses samples (passages and main ideas) and sources (already generated passages) to output main ideas	Main ideas of alternative passages
Title the Passage	GPT-3 uses samples (passages with titles) to output similar passages each with a title	Titles from alternative passages

ultimately be included in the subscores for comprehension and literacy by demonstrating (1) their association with other reading measures and (2) their psychometric qualities.

2.2.5 Test-Taker Readiness Materials and Practice Tests The Duolingo English Test delivers updates about any changes to the test that would impact the test taker experience well ahead in advance. In addition, an extensive readiness guide and unlimited practice tests are available at no cost to ensure that test takers' performance are free of bias due to unfamiliarity of test tasks and response format. The Duolingo English Test also communicates updates on the test via various media channels including YouTube, Facebook, and Twitter.

Table 3. Criteria for selecting correct options and distractors

Criteria for Selecting Correct Options	Criteria for Selecting Distractors
1. Higher estimated probability of generating that candidate 2. Similarity to the passage and individual sentences in the passage 3. Average similarity to other correct candidates 4. Other automated assessments by other machine learning models	1. Lower estimated probability of generating that candidate 2. Similarity to the passage and individual sentences in the passage 3. Average similarity to other correct candidates 4. Differences in likelihood and passage similarity from the selected correct candidate 5. Average similarity to other selected distractors 6. Other automated assessments by other machine learning models

3 Discussion

The Language Assessment Design Framework outlined in this paper helps build evidence for the digitally-informed chain of inferences for using Duolingo English Test scores for their intended purposes (Burstein et al., 2022), particularly pertaining to domain descriptions and scoring. Section 2.2.2 demonstrated how the different task types in Interactive Reading embody the constructs outlined in Section 2.2.1. The digital-first nature of the Duolingo English Test, including the mode of delivery for the input and the response format, helps fulfill the digital consideration of authentic and construct-relevant interaction on the test. The automated item generation methods allow Interactive Reading to administer texts that are representative of the TLU domain; the evidence specification activity shows that responses from Interactive Reading reflect reading skills.

This paper introduced a new item type on the Duolingo English Test called Interactive Reading that assesses reading comprehension and situated it within the Assessment Design Framework of the Assessment Ecosystem, drawing from theories on second language assessment and L2 reading. Interactive Reading strengthens the Duolingo English Test's validity claim of assessing reading in an academic context.

The addition of Interactive Reading is reflective of the digital-first nature of the Duolingo English Test in that Interactive Reading promotes efficiency and effectiveness at the same time. Interactive Reading is cost-efficient in that it utilizes state-of-the-art automatic item generation capabilities to generate a large number of reading passages and the accompanying items without involving extensive labor at the text and item generation stage, and it is effective in that the task is grounded in theories of second language (L2) reading and language assessment (Attali & von Davier, 2021). Interactive Reading is also part of a larger and continued effort of the Duolingo

English Test to expand its construct coverage since its inception, the examples of which include the addition of open-ended speaking and writing tasks (LaFlair, 2020) and the scoring of writing samples (Goodwin et al., 2022).

Not only is Interactive Reading an essential addition to the Duolingo English Test but it is also groundbreaking for the field of language assessment in that all passages, items, and options are generated automatically. Interactive Reading provides a potential spring board for further innovation on test development in that the format is flexible enough to support the development and inclusion of additional task types and different response formats. The format of digital assessment also provides the Duolingo English Test with the potential to evolve beyond an assessment tool and expand into a learning tool where feedback could be implemented as part of the assessment. Interactive Reading has already initiated the first step with the gradual reveal of the passage where rather than presenting a passage instantly, a passage is sequentially revealed across three different task types. Concurrent to this is the delivery of a subtle form of corrective feedback, whereby along with the rest of the passage, the answers to the previous tasks are also revealed. This provides a learning opportunity for test takers during the test as feedback, especially implicit feedback, is considered to be an effective learning tool (Li, 2010).

The last step in the digital chain of inferences is that the use of test scores is beneficial for all stakeholders involved. We argue that the inclusion of Interactive Reading in the scoring of the test is beneficial to score users as a digitally aligned, valid reading measure that will contribute to lowering barriers to education access, all the while maintaining a delightful test taker experience.

4 References

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732935

Attali, Y., & von Davier, A. (2021). *Computational psychometrics for test development* [Conference presentation].

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

Bell, K. E., & Limber, J. E. (2009). Reading skill, textbook marking, and course performance. *Literacy Research and Instruction*, *49*(1), 56–67. https://doi.org/10.1080/19388070802695879

Blanchard, J., & Mikkelson, V. (1987). Underlining performance outcomes in expository text. *The Journal of Educational Research*, *80*(4), 197–201. https://doi.org/10.1080/00220671.1987.10885751

Brantmeier, C. (2005). Effects of reader's knowledge, text type, and test type on L1 and L2 reading comprehension in Spanish. *The Modern Language Journal*, *89*(1), 37–53. https://doi.org/10.1111/j.0026-7902.2005.00264.x

Britt, M. A., Rouet, J.-F., & Durik, A. M. (2018). *Literacy beyond text comprehension: A theory of purposeful reading.* Routledge.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Burstein, J., LaFlair, G. T., Kunnan, A. J., & von Davier, A. (2022). *A theoretical assessment ecosystem for a digital-first assessment: The Duolingo English Test* (Duolingo Research Report DR-22-03 DRR-22-01). Duolingo. https://go.duolingo.com/ecosystem

Cardwell, R., LaFlair, G. T., & Settles, B. (2022). *Duolingo English Test: Technical manual*. Duolingo. https://duolingo-papers.s3.amazonaws.com/other/det-technical-manual-current.pdf

Cataldo, M. G., & Oakhill, J. (2000). Why are poor comprehenders inefficient searchers? An investigation into the effects of text representation and spatial memory on the ability to locate information in text. *Journal of Educational Psychology*, *92*(4), 791–799. https://doi.org/10.1037/0022-0663.92.4.791

Chapelle, C. (1999). From reading theory to testing practice. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency* (pp. 150–166). Cambridge University Press.

Cheng, J., & Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing*, *35*(1), 3–25. https://doi.org/10.1177/0265532216676851

Clapham, C. (1998). The effect of language proficiency and background knowledge on EAP students' reading comprehension. In A. J. Kunnan (Ed.), *Validation in Language Assessment* (pp. 141–168). Routledge. https://doi.org/10.4324/9780203053768

de Chazal, E. (2014). *English for academic purposes*. Oxford University Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper* (RM-00-4). Educational Testing Service. https://www.ets.org/Media/Research/pdf/RM-00-04.pdf

Eskey, D. E. (2005). Reading in a second language. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 563–580). Routledge.

Freedle, R., & Kostin, I. (1994). Can multiple-choice reading tests be construct-valid? A reply to Katz, Lautenschlager, Blackburn, and Harris. *Psychological Science*, *5*(2), 107–110. https://doi.org/10.1111/j.1467-9280.1994.tb00640.x

Goodwin, S., Attali, Y., LaFlair, G. T., Park, Y., Runge, A., von Davier, A., & Yancey, K. (2022). *Duolingo English Test–Writing construct* [Duolingo Research Report DR-22-03]. Duolingo. https://go.duolingo.com/scoredwriting

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge University Press. https://doi.org/10.1017/CBO9781139150484

Grabe, W., & Jiang, X. (2014). Assessing reading. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 185–200). John Wiley & Sons. https://doi.org/10.1002/9781118411360.wbcla060

Grabe, W., & Stoller, F. L. (2020). *Teaching and researching: Reading* (3rd ed.). Routledge.

Green, A., Ünaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing*, *27*(2), 191–211. https://doi.org/10.1177/0265532209349471

Guthrie, J. T. (1988). Locating information in documents: Examination of a cognitive model. *Reading Research Quarterly*, *23*(2), 178. https://doi.org/10.2307/747801

Guthrie, J. T., & Kirsch, I. S. (1987). Distinctions between reading comprehension and locating information in text. *Journal of Educational Psychology*, *79*(3), 220–227. https://doi.org/10.1037/0022-0663.79.3.220

Guthrie, J. T., & Mosenthal, P. (1987). Literacy as multidimensional: Locating information and reading comprehension. *Educational Psychologist*, *22*(3-4), 279–297. https://doi.org/10.1080/00461520.1987.9653053

Head, A. J., & Eisenberg, M. B. (2009). *Lessons learned: How college students seek information in the digital age* (No. 2). http://www.ssrn.com/abstract=2281478

Juffs, A. (2001). Psycholinguistically oriented second language research. *Annual Review of Applied Linguistics*, *21*, 207–220. https://doi.org/10.1017/S0267190501000125

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219–224. https://doi.org/10.3758/BF03194055

Krekeler, C. (2006). Language for special academic purposes (LSAP) testing: The effect of background knowledge revisited. *Language Testing*, *23*(1), 99–130. https://doi.org/10.1191/0265532206lt323oa

LaFlair, G. T. (2020). *Duolingo English Test: Subscores* (Duolingo Research Report DRR-20-03 DRR-20-03). https://duolingo-papers.s3.amazonaws.com/reports/subscore-whitepaper.pdf

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*(2), 309–365. https://doi.org/10.1111/j.1467-9922.2010.00561.x

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindquist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge, and use: New perspectives on assessment and corpus analysis* (pp. 57–78). European Second Language Association.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*(3), 513–536. https://doi.org/10.1111/1467-9922.00193

Qian, D. D., & Pan, M. (2014). Response formats. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 860–875). John Wiley & Sons. https://doi.org/10.1002/9781118411360.wbcla090

Qian, D. D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, *21*(1), 28–52. https://doi.org/10.1191/0265532204lt273oa

Rice, G. E. (1994). Examining constructs in reading comprehension using two presentation modes: Paper vs. computer. *Journal of Educational Computing Research*, *11*(2), 153–178. https://doi.org/10.2190/MV46-VW49-4G5W-FM92

Riley, G. L., & Lee, J. F. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. *Language Testing*, *13*(2), 173–189. https://doi.org/10.1177/026553229601300203

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

The Council of Europe. (2020). *Common European Framework of Reference for languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4.

Thompson, C., Morton, J., & Storch, N. (2013). Where from, who, why and how? A study of the use of sources by first year L2 university students. *Journal of English for Academic Purposes*, *12*(2), 99–109. https://doi.org/10.1016/j.jeap.2012.11.004

Urquhart, A. H., & Weir, C. J. (1998). *Reading in a second language: Process, product, and practice*. Routledge.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, *29*(3), 325–344. https://doi.org/10.1177/0265532211424478

Ward, W. C., Dupree, D., & Carlson, S. B. (1987). *A comparison of free-response and multiple-choice questions in the assessment of reading comprehension* (RR-87-20). Educational Testing Service. https://onlinelibrary.wiley.com/doi/10.1002/j.2330-8516.1987.tb00224.x

Weir, C., Hawkey, R., Green, A., Unaldi, A., & Devi, S. (2009). *The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university* (No. 9; pp. 97–156). IELTS Australia, British Council.

Winchell, A., Lan, A., & Mozer, M. (2020). Highlights as an early predictor of student comprehension and interests. *Cognitive Science*, *44*(11). https://doi.org/10.1111/cogs.12901

Yue, C. L., Storm, B. C., Kornell, N., & Bjork, E. L. (2015). Highlighting and its relation to distributed study and students' metacognitive beliefs. *Educational Psychology Review*, *27*(1), 69–78. https://doi.org/10.1007/s10648-014-9277-z

Zumbo, B. D., & Hubley, A. M. (2017). *Understanding and investigating response processes in validation research*. Springer.