

Sarah Goodwin*, Yigal Attali*, Geoffrey T. LaFlair*, Yena Park*, Andrew Runge*, Alina A. von Davier*, Kevin P. Yancey*, Ben Naismith*, Ping-Lin Chuang*

Abstract

Assessments, especially those used for high-stakes decision making, draw on evidence-based frameworks. Such frameworks inform every aspect of the testing process, from development to results reporting. The frameworks that language assessment professionals use draw on theory in language learning, assessment design, and measurement and psychometrics in order to provide underpinnings for the evaluation of language skills including speaking, writing, reading, and listening. This paper focuses on the construct, or underlying trait, of writing ability. The paper conceptualizes the writing construct for the Duolingo English Test (DET). This work is situated in the Burstein et al. (2022) theoretical ecosystem for digital-first assessment, the first representation of its kind that incorporates design, validation/measurement, and security all situated directly in assessment practices that are digital first. The paper first provides background information about the DET and then defines the writing construct, including the purposes for writing. It also introduces principles underpinning the design of writing items and illustrates sample items that assess the writing construct. Furthermore, it details the DET’s integration of dynamic task types and the application of artificial intelligence (AI) in item generation and real-time adaptation, presenting a distinctive approach to evaluating the writing construct in a computer-based testing environment.

Last Update: June 2025

Keywords

Duolingo English Test, digital-first assessment, language assessment, L2 writing, second language writing, writing assessment

Contents

1	Introduction	3
2	Theoretical background for DET writing assessment	3
2.1	Cognitive processes and writing skills	3
2.2	CEFR written production	4
2.3	Purposes of writing	4
2.4	Sociocognitive frameworks	5
3	Item design	5
3.1	Picture Description (Writing)	5
3.2	Interactive Listening Summarization	6
3.3	Interactive Writing	6
3.4	Writing Sample	6
4	Item development and scoring	8
5	Evidence-specification activity	8
6	Test-taker readiness materials and practice tests	8

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

7	Summary	8
8	References	10

1 Introduction

The Duolingo English Test is a digital-first*, computerized-adaptive, high-stakes English language proficiency assessment. A primary test purpose is to inform admissions decisions at English-medium universities (Naismith et al., 2025). The DET assesses the four language constructs of Speaking, Writing, Reading, and Listening (SWRL), and it is aligned to the proficiency levels and can-do statements for the Common European Framework of Reference (CEFR) for Languages (Council of Europe, 2001, 2020). SWRL skills are important for success in English-medium universities, so it is important that items are designed to assess the knowledge, skills, and abilities necessary in university English language tasks. Writing is assessed with four task types: Picture Description (writing), Interactive Listening Summarization, Interactive Writing, and the Writing Sample. In 2022, the decision was made to score the previously-unscored Writing Sample to increase the reliability of measurement and improve construct coverage of writing, while maintaining the test length. Interactive Listening Summarization was added to the DET in 2023, and Interactive Writing replaced Read Then Write on the official test in April 2024.

The DET contains a variety of task types that contribute to nine different scores: an overall scaled score; four independent subscores for Speaking, Writing, Reading, and Listening; and four integrated subscores Literacy, Conversation, Comprehension, and Production (Naismith et al., 2025). The test employs human-in-the-loop† AI, using technology with humans involved throughout every step of item development, automated scoring of responses, and proctoring and security procedures. It is designed to assess English proficiency while lowering barriers to test access and providing a positive test taker experience. The DET does this in the following ways: 1) the test is secure and can be taken any time remotely at home; 2) it is a computerized-adaptive test, with a shorter time commitment than many other standardized tests; 3) it has a lower price point than other high-stakes language assessments on the market; and 4) its free test readiness guide and practice tests promote greater access.

2 Theoretical background for DET writing assessment

The subsections below provide an overview of the theoretical foundations of the writing tasks on the DET and situate them within Burstein et al. (2022)’s ecosystem in the Language Assessment Design Framework. The Language Assessment Design Framework includes construct definition, sociocognitive frameworks (Mislevy, 2018; Weir, 2005), and natural language processing (NLP) and AI for automated item generation, scoring, and item difficulty prediction. These factors are implemented with attention to language assessment theory. The ecosystem incorporates foundational test validity argument theory (Chapelle et al., 2011; Kane, 1992, 2010) to support test score interpretation and use. The sections below include more detail about construct definition, item design, automated item generation and scoring, evidence-specification activities (how decisions are made from test taker data, and the specifications of data pipelines), and test-taker readiness materials.

2.1 Cognitive processes and writing skills

Writing can be categorized by hierarchical levels of cognitive processing (Vähäpääsi, 1982; Weigle, 2002):

1. To reproduce (predetermined information)
2. To organize/reorganize (known spatial/temporal phenomena/concepts/mental states)
3. To invent/generate (new or alternative spatial/temporal phenomena/concepts/mental states)

Examples of writing at hierarchical level 1, reproduction, include filling in a form or copying down speech verbatim. Examples at the second level, requiring organization or reorganization, involve arranging information that is known to the writer, such as writing a lab report, producing a summary, crafting an outline, and/or paraphrasing. The most demanding type of cognitive processing is the third level: writing for knowledge transformation and idea generation, characteristic of university settings.

As writers build their discipline-specific writing skills, for instance in postsecondary education, they move from learning-to-write to writing-to-learn (Manchón, 2011). Writers who are not writing in the language they are most familiar with may still be in the learning-to-write phase, learning some components of writing for general purposes, while also acquiring the language of their discipline. More proficient writers compose differently from less proficient writers (Bereiter & Scardamalia, 1987; Hyland, 2022; Silva, 1993).

*Digital-first assessment includes technology such as artificial intelligence (AI) and machine learning, with human expert involvement, throughout all item development, test scoring, and security processes.

†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).

While many of these developmental patterns are observed across writers in general, additional considerations emerge when writers are composing in a language that is not their first. Second language (L2) writers, although a heterogeneous group, do tend to have some differences from first language (L1) writers, who also differ a great deal. “Skill”, as used below, can be thought of as the degree to which writers can effectively use their language knowledge as well as draw on strategic competencies. Language knowledge and strategic competence are interrelated components of language ability (Bachman & Palmer, 2010; Swain, 1985) for language learners to show their writing ability:

- Composing processes are similar in first and additional languages
- More skilled writers compose, plan, and revise more effectively than novice writers
- Able writers focus on revisions more at the global level, while less-skilled writers revise at the word and/or sentence level
- Strategies in L1 contexts will not necessarily transfer to L2 contexts
- L2 writers tend to plan less than L1 writers
- L2 writers, compared to L1, have more trouble goal-setting and generating material
- L2 writers, compared to L1, revise more, reflect less (Hyland, 2022; Silva, 1993)

There is much in the L2 writing literature that focuses on the writing process. However, it is important to note that any timed impromptu writing assessment administered at just one time point measures only a snapshot of writing. The relatively short task duration does not permit much time for someone to create more than one draft and revise their written work. An alternative to impromptu writing is portfolio assessment, in which writers submit multiple drafts of their writing, or the versions of writing that are a best reflection of their improvement or achievement, and portfolios may include reflections on the writing process (Hamp-Lyons & Condon, 1993; Weigle, 2002). A writing assessment that is given at only one point in time, thus, is limited in how much it can reflect test takers’ complex competencies.

Users of English at advanced levels, including students in university settings, must be able to show their ability to write texts for a variety of audiences and purposes, including explaining information and expressing opinions arguing for or against points of view. Test takers may not necessarily be pursuing goals that require them to show their English production skills for university settings. Hence, DET task design permits learners at lower proficiency levels to be able to produce samples of extended writing, while still also allowing more advanced English writers to respond with a more complex written performance. In order to accomplish this, Duolingo English Test writing tasks are designed and scored in ways that support broad generalizations about test-taker writing ability, which includes writing in university settings.

2.2 CEFR written production

The writing construct draws on salient characteristics from the Common European Framework of Reference (CEFR) for Languages (Council of Europe, 2001, 2009, 2020). The CEFR provides can-do statements about language ability, classifying language use into public, personal, educational, and/or occupational domains. Domains can be thought of as the contexts language use takes place in. The same type of writing does not always fall into the same domain; for instance, an email or direct-messaging conversation could occur in the personal domain (between friends) or in the educational domain (between students, or a student and a professor). In addition, public and personal domains do not necessarily contain less formal types of language use than educational and occupational ones.

The CEFR describes language use ranging from A1 (basic user) to C2 (proficient user). Writing activities at A1 and A2 levels are mainly personal and/or have a formulaic structure such as a simple letter to a friend or family member. B1 and B2 levels move from the personal to more information-delivery types of texts found in public and occupational spheres. Test takers at C1 and C2 levels can produce sophisticated writing, characterized by strong cohesion and organization and a range of accurate vocabulary, grammar, and idiomatic phrases. The DET aims to assess writing ability across all CEFR levels, assessing written production appropriate to the various CEFR domains. DET writing tasks elicit responses that are evaluated using well-defined subconstructs: Content, Discourse Coherence, Grammar, and Vocabulary. These subconstructs are grounded in CEFR descriptors and operationalized through both human and automated scoring systems. They provide targeted evidence of writing proficiency aligned with the CEFR framework, ensuring that the DET writing scores meaningfully reflect test takers’ ability to communicate in written English across real-world contexts and levels of complexity.

2.3 Purposes of writing

Dominant purposes for writing include but are not limited to: writing to learn; to convey emotions or feelings; to inform; to convince or persuade; to entertain, delight, or please; and to keep in touch (Vähäpääsi, 1982; Weigle, 2002). Note that there is not an implied hierarchy among these six purposes. Vähäpääsi (1982) also considered whether the audience for the writing was oneself, others, or both. Other key purposes commonly found in writing assessment are writing for narration, description, and argumentation. Specifically, in university settings, more so than during students’ formative years or in workplace writing, college writing tasks involve being able to critically read and analyze text, such as writing annotated bibliographies or research documents (Burstein et al., 2016).

The CEFR companion volume (Council of Europe, 2020) differentiates the subskills of writing production and interaction by conceiving of Production as being “sustained monologue” characterized by “long turns” and Interaction as “conversational dialogue” consisting of “short turns” (Council of Europe, 2020, p. 34). Production and interaction are both components of writing ability. Written Production involves the writing of, from the C2 descriptor, “clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader identify significant points” (Council of Europe, 2020, p. 34). Written Interaction involves interactive communication, including interpersonal exchanges. The CEFR conceptualizes online communication separately because it is multimodal (Council of Europe, 2020). The writing construct thus involves written production and interaction ability as well as several interrelated communicative strategies, including language users’ abilities to plan, compensate, monitor, self-correct, take turns, cooperate, repair communication breakdowns, and clarify. The CEFR recognizes that these planning and compensation strategies are an important part of language constructs, especially as users move from pre-A1 to eventually B2, C1, or C2 performance levels. Sociocognitive frameworks (Mislevy, 2018; Weir, 2005) give us a way to conceptualize these as secondary factors with a primary factor being the assessment of the given construct: in this case, writing.

2.4 Sociocognitive frameworks

In addition to the description of cognitive underpinnings of and purposes for writing, the DET ecosystem framework utilizes the sociocognitive framework (Mislevy, 2018; Weir, 2005). Table 1 describes the task constructs, both primary and secondary, including general skills and influential factors that will always play roles in test taker performance:

Table 1. Primary and Secondary Constructs and Related Factors in Writing

Primary construct	Writing
Secondary constructs	1. Reading. On a more-independent to more integrated writing task continuum, DET writing tasks would fall more toward the independent end, with fewer extensive reading source text(s) 2. Pragmatic skills (e.g., politeness) 3. Interactional skills (e.g., collaboration, turn taking)
General skills	1. Critical thinking 2. Content knowledge
Other influential factors	1. Intrapersonal factors (e.g., confidence, motivation) 2. Experiential factors 3. Neurological/physiological factors

It is critical to note that skills in addition to writing, as well as related experiences, contribute to successful language use. Sociocognitive frameworks, for example, take into account intrapersonal factors like confidence, as well as experiential and physiological variables. An example of an experiential factor is a test taker’s familiarity with a specific test format, which can affect performance. Likewise, physiological factors, like the ability to sustain focus over long periods, can have an impact. Shorter writing tasks can have test-retest reliability and criterion validity that is just as strong as, or stronger than, longer tasks (Naismith et al., 2024).

Acknowledging these variables is key to minimizing construct-irrelevant variance in DET items. To that end, test takers can use the freely available Practice Test to familiarize themselves with the question format. They also participate in experimental test sections, which allows our developers to analyze results and make informed decisions about timing and user experience.

3 Item design

On a DET administration, test takers respond to four types of writing tasks: Picture Description (writing), Interactive Listening Summarization, Interactive Writing, and the Writing Sample. For more information about these task types, see Naismith et al. (2025), Goodwin et al. (2024), LaFlair et al. (2025).

3.1 Picture Description (Writing)

The Picture Description (writing) task requires test takers to describe an image (see Figure 1). This task assesses the ability to use varied vocabulary, construct grammatically correct sentences, maintain coherence, generate descriptive language, and translate visual observations into written text.

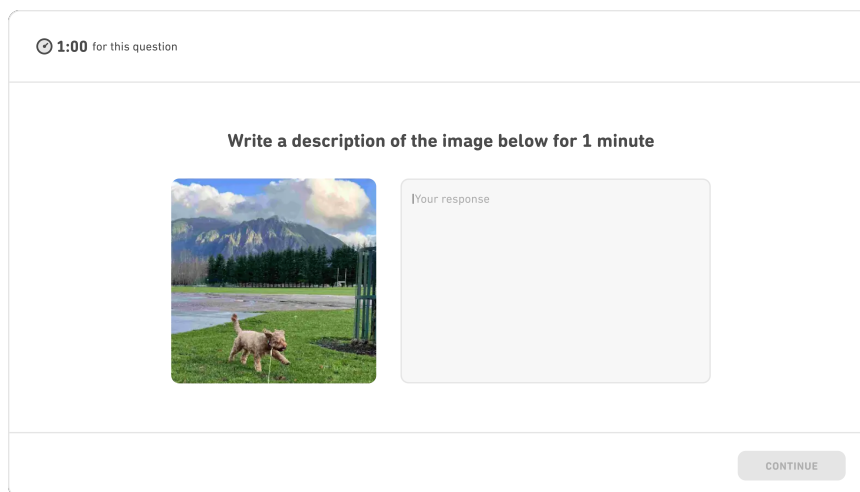


Figure 1. A screenshot of the Picture Description task on the Duolingo English Test

3.2 Interactive Listening Summarization

The Interactive Listening (Summarization component) task (see Figure 2), following an interactive listening scenario and conversation, requires test takers to compose a written summary of the conversation. This component taps into the ability to comprehend spoken English, extract key information, synthesize it, and present it coherently and concisely in writing.

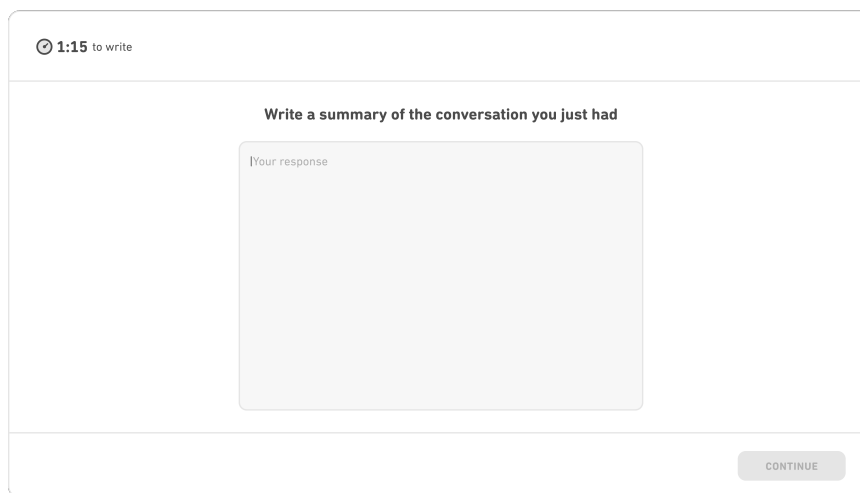


Figure 2. A screenshot of the Interactive Listening task on the Duolingo English Test - summarization

3.3 Interactive Writing

Interactive Writing is a two-part task. Initially, test takers respond to a prompt (see Figure 3), and subsequently, the system presents a follow-up question based on their initial input (see Figure 4). This task is designed to assess the ability to produce cohesive and coherent discourse and adapt their writing based on follow-up prompting.

3.4 Writing Sample

Finally, the Writing Sample task (see Figure 5) requires test takers to produce an extended piece of writing in response to a given prompt. This allows for an assessment of their ability to construct structured and developed persuasive or expository texts. The full text of this Writing Sample is provided to institutions. Since 2022, the task has contributed to the DET score, to increase measurement reliability and construct coverage.

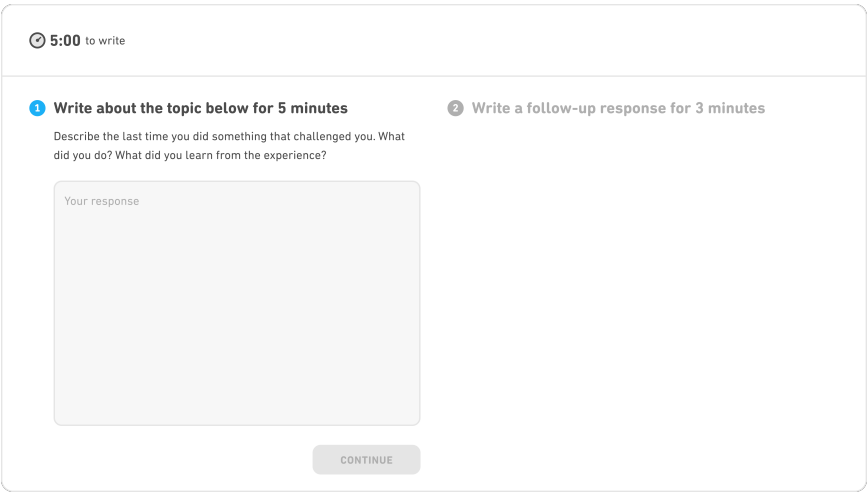


Figure 3. A screenshot of the Interactive Writing task on the Duolingo English Test - initial prompt

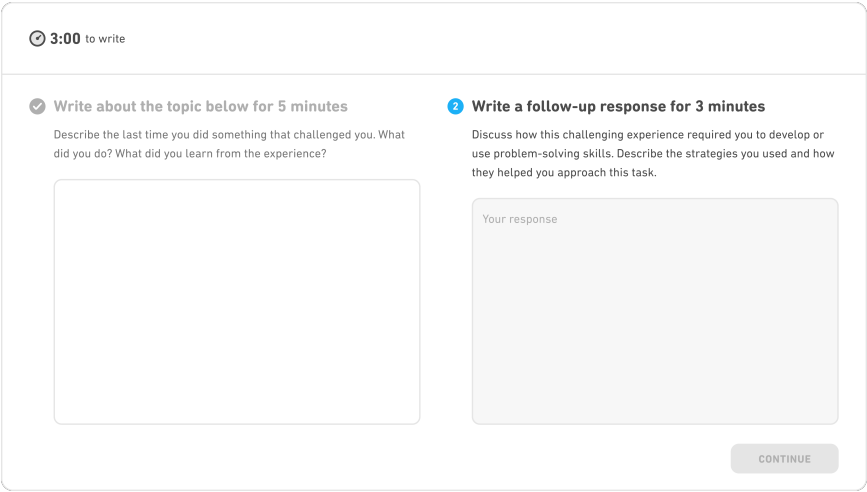


Figure 4. A screenshot of the Interactive Writing task on the Duolingo English Test - follow-up prompt

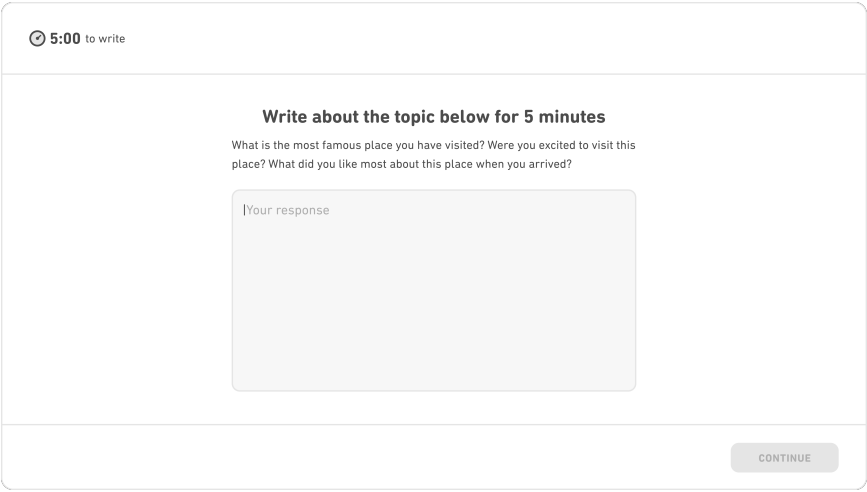


Figure 5. A screenshot of the Writing Sample task on the Duolingo English Test

4 Item development and scoring

The test design process is a collaborative effort that involves multiple layers of expertise. Test design approval is made by not just assessment research staff but also with input from executive-level design staff and product managers. In addition, Duolingo conducts user experience design research with people who live all over the world, with varying computer skills and various speeds of internet connectivity. Items' text, audio, and images are tested so that they load quickly under acceptable standards. These steps are followed to check that test takers can access and interact with the items as test item developers expect, and to meet legal and industry standards.

For text excerpts, some prompts are written by expert test developers. Other prompts are generated using the large language model GPT (Generative Pretrained Transformer; [OpenAI, 2025](#)). The text outputs from automated item generation are reviewed for appropriateness of topics and text types by language assessment experts as part of the human-in-the-loop AI procedures. The image sources comprise a human-curated collection of stock photographs from open sources. All stimuli, including prompts, passages, item options, and images, are reviewed by subject-matter experts experienced in test development. All items also undergo fairness and bias review by experts in language learning, assessment, and intercultural communication.

Automated scoring features include measures that assess content, organization, and accuracy and range of grammar and vocabulary. Given that machine grading is fundamentally different from human grading ([Attali, 2013](#)), the automated scoring approach balances the predictive validity and reliability of grades. Candidates should produce a well-developed, on-topic response that employs appropriate grammar and vocabulary. Writers should ensure that they are communicating their message clearly while also showing all of the strengths of their English writing skills they can.

5 Evidence-specification activity

The evidence specification stage concerns how well the scores will represent the Duolingo English Test writing construct. For the writing item types, continuous response data are used to contribute to test taker scores. The data for the evidence specification phase come from certified test repeat administrations during the 2020 calendar year (N=47,640). The relationship between the three writing tasks was compared, to justify changing the Writing Sample to be a scored component without increasing total test time. All writing task types were moderately to highly correlated with each other, with strong relationships between the extended samples of writing Read Then Write and Writing Sample. Data from retesters revealed the Writing Sample had a higher test-retest reliability than the Write About the Photo task. Adding the third sample as a scored writing sample also increased test-retest reliability of the aggregate writing score. In addition, the scores for the writing tasks are invariant by country and first language (L1) background and correlate with other measures of English proficiency. Duolingo English Test writing items are designed to support data collection relevant to score use for university admissions purposes. Automated scoring features can be analyzed in the future and contribute to systems of intelligent-tutor writing feedback for learners. These attributes tie into the Burstein et al. (2022) ecosystem to inform test result interpretation and use. Drawing on this scaffolding of multiple theoretical frameworks means research inquiry can be approached in a systematic way, in order to inform item development and psychometric processes.

6 Test-taker readiness materials and practice tests

The certified version of the Duolingo English Test is taken in a downloaded, executable computer program. To prepare, test takers can access a full-length practice test that simulates the adaptive nature and item presentation of the certified test, or they can access certain sections in the Practice Hub. Practice test takers must have a desktop or laptop computer with a camera and microphone, and allow the application access to those. The administration of the practice test is similar to the certified test.

Test takers also can meet with and ask questions of our Institutional Engagement staff via Duolingo English Test webinars and livestreams. Often in partnership with accepting institutions, many of these video streams describe the format and sections of the test. Our social media channels have approachable overviews of the test sections to increase accessibility and test access.

7 Summary

The paper described the construct of writing, situating it in Burstein et al. (2022)'s ecosystem of validation, design, measurement, and security all situated in theoretical frameworks for assessment practices that are digital first. The writing construct is informed by frameworks such as the CEFR and sociocognitive framework, as well as writing cognitive processes documented in the literature. The Duolingo English Test has a variety of tasks to assess writing: picture description, summarization, and extended writing with elaboration. Along with test scores and subscores provided to institutions, the full text of the Writing Sample allows test takers to offer a window into their writing ability that they can show to admissions

officials. This work connects to the Duolingo English Test mission to use technology to lower barriers to assessment and increase opportunities for English language learners everywhere.

Acknowledgements

We thank our colleagues Jill Burstein, Ramsey Cardwell, Antony John Kunnan, Mya Poe, and Xiaowan Zhang, for their assistance with writing task development and feedback on this paper.

8 References

- Association of Test Publishers. (2024). *Creating responsible and ethical AI policies for assessment organizations*.
- Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181–198). Routledge.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*.
- Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Lawrence Erlbaum Associates.
- Burstein, J., Elliot, N., & Molloy, H. (2016). Informing automated writing evaluation using the lens of genre: Two studies. *CALICO Journal*, 33(1), 117–141.
- Burstein, J., LaFlair, G. T., Kunnan, A. J., & Davier, A. A. (2022). *A theoretical assessment ecosystem for a digital-first assessment* [Duolingo Research Report DRR-22-01]. <https://go.duolingo.com/ecosystem>
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2011). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.
- Council of Europe. (2001). *Common European Framework of Reference for Languages*. <https://rm.coe.int/1680459f97>
- Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages*. <https://www.coe.int/en/web/common-european-framework-reference-languages/relating-examinations-to-the-cefr>
- Council of Europe. (2020). *Common european framework of reference for languages*. <https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4>
- Goodwin, S., Poe, M., Cardwell, R., Runge, A., Attali, Y., Mulcaire, P., Lo, K.-L., & LaFlair, G. T. (2024). *Facilitating the writing process on the DET: The interactive writing task* [White Paper]. Duolingo English Test. <https://duolingo-papers.s3.us-east-1.amazonaws.com/other/interactive-writing-whitepaper.pdf>
- Hamp-Lyons, L., & Condon, W. (1993). Questioning assumptions about portfolio-based assessment. *College Composition and Communication*, 44(2), 176–190.
- Hyland, K. (2022). *Teaching and researching writing* (4th ed.). Routledge.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527–535.
- Kane, M. T. (2010). Validity and fairness. *Language Testing*, 27(2), 177–182.
- LaFlair, G. T., Runge, A., Attali, Y., Park, Y., Church, J., & Goodwin, S. (2025). *Interactive listening—The Duolingo English Test* (Duolingo Research Report DRR-23-01). Duolingo.
- Manchón, R. (Ed.). (2011). *Learning-to-write and writing-to-learn in an additional language*. John Benjamins.
- Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.
- Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, A. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56, 3005–3054. <https://doi.org/10.1007/s10462-022-10246-w>
- Munro, R. (2021). *Human-in-the-loop machine learning: Active learning and annotation for human-centered AI*. Manning Publications.
- Naismith, B., Attali, Y., & LaFlair, G. T. (2024). The impact of task duration on the scoring of independent writing responses of adult L2-english writers. *Assessing Writing*, 62, 100895. <https://doi.org/10.1016/j.asw.2024.100895>
- Naismith, B., Cardwell, R., LaFlair, G., Nydick, S., & Kostromitina, M. (2025). *Duolingo English Test: Technical manual*. Duolingo. <https://go.duolingo.com/dettechnicalmanual>
- OpenAI. (2025). *GPT family models (GPT-3.5, GPT-4, GPT-4o) [large language models]*. <https://platform.openai.com>.
- Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly*, 27(4), 657–677.
- Swain, M. (1985). Large-scale communicative language testing: A case study. In Y. P. Lee, A. C. Y. Fok, R. Lord, & G. Low (Eds.), *New directions in language testing* (pp. 35–46).
- Vähäpääsi, A. (1982). On the specification of the domain of school writing. In A. C. Purves & S. Takala (Eds.), *An international perspective on the evaluation of written composition* (pp. 265–289).
- Wang, G. (2019). *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.
- Weir, C. J. (2005). *Language testing and validation*. Palgrave MacMillan.