A Theoretical Assessment Ecosystem for a Digital-First Assessment—The Duolingo English Test

duolingo english test

Duolingo Research Report DRR-21-04
August 24, 2021 (32 pages)
englishtest.duolingo.com/research

Jill Burstein*, Geoffrey T. LaFlair*, Antony John Kunnan*, Alina A. von Davier*

Abstract

The *Duolingo English Test* is a groundbreaking, digital-first, computer-adaptive English language proficiency test. In contrast to traditional assessments, the *Duolingo English Test* is a digital-first assessment, leveraging "human-in-the-loop AI" from end to end to automatically generate test items and score test-taker item responses. The test is aligned with the Common European Framework of Reference for Languages (CEFR) English proficiency levels, and leverages a sociocognitive approach to inform test item construct definition. As digital-first assessments are the future, there needs to be a comprehensive framing for these assessments. Yet, to our knowledge, no such framing is available. This paper presents a novel theoretical assessment ecosystem for the *Duolingo English Test*. The ecosystem formalizes the theory underlying the test and key processes that support construct validity and a chain of inferences. The chain of inferences addresses digital considerations to ensure a valid, fair and reliable test score that can be used for the intended purpose of informing stakeholder admissions decisions at English-medium institutions. The ecosystem is composed of a coherent, comprehensive, and integrated set of complex frameworks that considers digitally-mediated test facets. The frameworks include: (1) the Language Assessment Design Framework, (2) the Expanded Evidence-Centered Design Framework, (3) the Computational Psychometrics Framework, and (4) the Test Security Framework. The expected impact of the test facilitates Duolingo's social mission to lower barriers to education access and offer a secure and delightful test experience, and to provide a valid, fair and reliable test score. The ecosystem leverages principles from assessment theory, computational psychometrics, design, data science, language assessment theory, NLP/AI and machine learning, and test security.

Keywords

Duolingo English Test, digital-first assessment, language assessment

Contents

*Duolingo, Inc.

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

1 The Duolingo English Test

The *Duolingo English Test* is a groundbreaking, digital-first, computer-adaptive English language proficiency test (Settles et al., 2020). The test assesses four key constructs for university English language proficiency: Speaking, Writing, Reading and Listening (SWRL), and is aligned with the Common European Framework of Reference for Languages (CEFR) proficiency levels (The Council of Europe, 2001, 2020). *Duolingo English Test* scores are intended to be used by stakeholders to inform admissions decisions at English-medium institutions. Test subscores include Comprehension, Conversation, Literacy, and Production; these subscores represent integrated language skills that offer a more nuanced evaluation of test-taker abilities (LaFlair, 2020). See Table A2 for a list of test item types and associated constructs and subscore categories. The test-taker experience (or, TTX) is a key consideration across the test from item design to test administration to institutional score reporting processes (von Davier, 2021). As a digital-first assessment, the test leverages "human-in-the-loop AI" from end to end. Specifically, AI is used for test security, and automated generation of test items and scoring of test-taker responses (Settles et al., 2020). Humans are involved in test proctoring processes, review of automatically-generated test items, and monitoring of automated scoring.

The test is tied to Duolingo's mission to promote positive social impact by lowering barriers to access the test and providing a positive TTX from end-to-end. Factors that support this goal include the following: (a) it is the first test to support 24/7, secure, remote, at-home testing, (b) it is a computer-adaptive test; thus, it is shorter than traditional assessments, (c) it is offered at an affordable cost ($49 USD), promoting wider test access, and (d) the test's website offers a free test readiness guide and practice tests, further increasing test access. Currently, the *Duolingo English Test* has been adopted for use at more than 3,500 programs in 70 countries, and its international test adoption continues to grow. Test score concordances with the IELTS* and TOEFL® iBT† assessments suggest that the *Duolingo English Test* is a comparable measure of English language proficiency.

1.1 Ecosystem Rationale

Different frameworks and guiding principles for language assessment design, and assessment design more generally, have been presented in the assessment literature. Yet, to our knowledge, there is no single framing that comprehensively embodies the complex set of assessment development and evaluation processes and interactions, and digital considerations essential for digital-first assessment. The *Duolingo English Test* ecosystem combines a comprehensive and integrated set of complex frameworks that guide key steps and decisions in assessment development and evaluation. Through the ecosystem, a chain of inferences is built that supports test score interpretation and use (Chapelle et al., 2008; Kane, 1992, 2011). The chain of inferences is digitally-informed; it supports the test's expected impact and drives *Duolingo English Test* innovation. In this paper, we use the term expected impact similarly to positive

*https://www.ielts.org/en-us
†https://www.ets.org/toefl

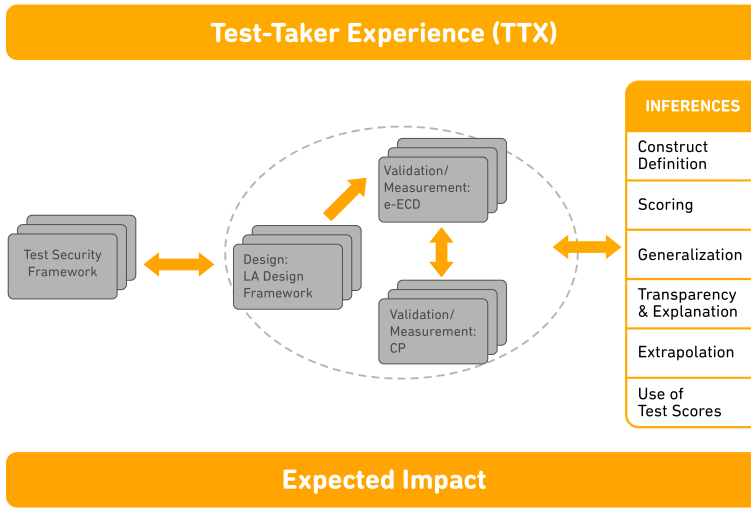

Figure 1. Illustration of the ecosystem framework components and interactions, and the connection to the digitally-informed chain of inferences (DCI). TTX influences all ecosystem components and the DCI. Expected Impact is satisfied as TTX and ecosystem and DCI decisions are implemented.

intended consequences (Kane, 2013; Messick, 1989). Specifically, the expected impact assumes the following outcome: to facilitate Duolingo's social mission to lower barriers to education access and offer a delightful test-taker experience, and to provide a valid, fair and reliable test score. Figure 1 illustrates the *Duolingo English Test* ecosystem and its interactions with the digitally-informed chain of inferences. Further, the figure shows that TTX affects decisions within and across ecosystem components. As a result, TTX can influence the digitally-informed chain of inferences. The figure indicates that Expected Impact is a result of decisions associated with TTX, the ecosystem, and the digitally-informed chain of inferences.

1.1.1 Existing Assessment Frameworks

Different assessment frameworks have been developed for varying assessment purposes and contexts, a few of which are described in this section. The CEFR offers a framework that provides levels and descriptors for teaching, learning and assessing language proficiency in Europe (The Council of Europe, 2001, 2020). English language proficiency assessments can use CEFR levels and descriptors to conceptualize and iterate on item type design. Mislevy et al. (2003)'s Evidence-Centered Design Framework provided a blueprint for a conceptual framework for educational assessments. The framework supports an evidentiary argument about student knowledge, skills, and abilities. It includes task design, test configuration, development of feature measures, and statistical modeling of relevant measures to generate a student model. Building on this, the Expanded Evidence-Centered Design Framework (e-ECD) adds a learning branch to the ECD framework that supports formative assessment and instruction (Arieli-Attali et al., 2019). See Section 2.2 for

more discussion of ECD and e-ECD. Papageorgiou et al. (2021) presented a framework for the TOEFL® Essentials™—a recent digital-first language assessment. They describe a traditional language assessment framework that underlies the test, including test and item design, item types, scoring and score interpretation, and security measures implemented for the test. Tannenbaum and Katz (2021) presented a framework that outlines validity considerations in the development of complex task design, taking digital performance tasks into account. Barrett et al. (2021) discussed a "smart authoring system" that illustrates the design, configuration and deployment of adaptive assessments, and outlined six principles supporting iteration and focussed attention on the user experience. Cope et al. (2020) proposed a framework that outlines "opportunities and boundaries" for AI in education. They examine different artifacts and processes that can be captured between traditional and AI-enabled assessments. For example, they illustrated the difference in the breadth of data types that can be collected from traditional (narrower range) versus digital (wider range) assessments. ATP (2021) advised that assessments that use AI should be guided by the following set of principles: (1) privacy; (2) accountability; (3) safety and security; (4) transparency and explainability; (5) fairness; (6) human control of technology (i.e., human in the loop); (7) professional responsibility (e.g., valid scores); and (8) promotion of human values (see Fjeld et al., 2020). Van Moere and Downey (2016) discussed the need to consider technology and AI as we build validity arguments for assessments. These individual frameworks and principles offer different perspectives on building and evaluating assessments. However, none of them combine and describe interactions across the full set of processes and considerations needed for the *Duolingo English Test*, especially with regard to necessary digital considerations. The following section provides a description of the *Duolingo English Test* ecosystem.

1.1.2 The Duolingo English Test Ecosystem

In contrast to existing frameworks and guiding principles, the *Duolingo English Test* ecosystem is composed of a coherent, comprehensive, and integrated set of complex assessment frameworks that addresses the prominent digital attributes of the test, and supports a novel, digitally-informed chain of inferences. The ecosystem frameworks include: (1) the Language Assessment Design Framework, (2) the Expanded Evidence-Centered Design (e-ECD) Framework, (3) the Computational Psychometrics Framework, and (4) the Test Security Framework. TTX spans the entire ecosystem (see Figure 1). TTX considerations include factors such as low price point and shorter testing time, free test-readiness resources, delightful UX design, accessibility and accommodations, and fast score turn-around processes.

Figure 1 illustrates the *Duolingo English Test* ecosystem and its relationship to the digitally-informed chain of inferences (Table A1). The ecosystem and the chain of inferences work hand-in-hand to achieve the test's expected impact with regard to the test's social mission and test score. Table A1 illustrates how the *Duolingo English Test* builds a digitally-informed chain of inferences, interacting with each of the ecosystem frameworks. For each inference, the table provides examples of digital considerations associated with the different ecosystem frameworks. As digital considerations within a framework are satisfied, the test gets closer to achieving the expected impact—i.e., to lower barriers to access, ensure a delightful test-taker experience, and provide valid, fair, and reliable test scores. This is consistent with Bachman and Palmer (2010),

Chalhoub-Deville (2009) and Chalhoub-Deville and O'Sullivan (2020) who assert the critical importance of a systematic process that ensures that the test yields the expected impact.

Chapelle et al. (2008) proposed a chain of inferences to support a validity argument for the TOEFL® assessment—a high-stakes test used for admissions to English-medium universities. They provided six inference types, each aligned with a warrant and a set of underlying assumptions for each inference. The warrant is an assumption tied to a claim about test score interpretation. For example, the claim might be that a test taker's score suggests that their university English language skills are sufficient to be successful at an institution. Chapelle et al. (2008)'s six inferences are adapted by the *Duolingo English Test* to build a novel, digitally-informed chain of inferences. A summary of the six inferences is provided here.

1. **Domain Description** is associated with the warrant that the *Duolingo English Test* item types represent knowledge, skills, and abilities associated with constructs relevant to university English language skills required for English-medium institutions, including digitally-mediated communication;
2. **Scoring** is associated with the warrant that observed *Duolingo English Test* performance based on automated evaluation methods is reflective of university English language skills required for English-medium institutions. It assumes, for example, that automatically-derived scoring feature measures are construct relevant, provide appropriate evidence of skills, and offer explanation for language proficiency outcomes;
3. **Generalization** is associated with the warrant that observed *Duolingo English Test* performance measures are estimates of expected performance for parallel versions of an automatically-generated test, and across automated and human raters and test administrations. An example assumption is that different task configurations will support the intended interpretation and are equitable;
4. **Transparency & Explanation** is related to the warrant that observed *Duolingo English Test* performance provides interpretable English language proficiency measures consistent with university English language skills required for English-medium institutions. It assumes that both computationally-derived feature measures used for scoring and evaluation and the internal structure of test scores are transparent and explainable, and are aligned with theoretical language proficiency attributes (i.e., construct attributes);
5. **Extrapolation** is associated with the warrant that the test assesses the construct of English language proficiency consistent with university English language skills required for English-medium language institutions. It assumes that observed test performance based on automated scoring outputs is related to relevant external measures of academic proficiency;
6. **Use of Test Scores** is related to the warrant that observed *Duolingo English Test* performance is beneficial for stakeholders. The inference assumes that automatically-generated feature measures and scores provide interpretable evidence of English language proficiency that support stakeholder decisions.

In contrast to Chapelle et al. (2008), the *Duolingo English Test*'s chain of inferences explicitly considers how to satisfy theoretical underlying assumptions given digital considerations across the ecosystem (see Table A1). Let's consider an example for the Transparency & Explanation inference. In this case, a digital consideration might be to evaluate if AI methods produce interpretable measures that can be clearly mapped to relevant university English language skills attributes. For example, for a test-taker's written response, does the AI produce interpretable measures associated with vocabulary usage quality? As digital considerations are addressed across the ecosystem, a digitally-informed chain of inferences is built that can support an explainable and defensible test score. Further, the test considers the "impact of technology" (p. 4) such as automated scoring of essays and complex, innovative item types (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014). Table A1 illustrates the digitally-informed chain of inferences. The table provides example assumptions that underlie the inferences and associated digital considerations across the ecosystem frameworks. The Duolingo English Test's novel, digitally-informed chain of inferences includes a broader set of digital considerations than those presented in Xi et al. (2008)'s and Xi (2010)'s. Xi et al. (2008)'s and Xi (2010)'s, respectively, focus only on digital considerations for automated speech and essay scoring used on assessments.

2 Ecosystem Overview

The *Duolingo English Test* ecosystem embodies the theory underlying the test that guides item design, decisions about evidence collection, and modeling approaches that inform test score interpretations and score use (Kane, 2013; Messick, 1995; Mislevy et al., 2003). Figure 1 illustrates the ecosystem framework components and interactions through which a digitally-informed chain of inferences is constructed: (1) the Language Assessment Framework, (2) the Expanded Evidence-Centered Design (e-ECD) Framework, (3) the Computational Psychometrics Framework, and (4) the Test Security Framework. Test-taker experience (TTX) spans the entire ecosystem. We distinguish TTX from UX (user experience). While UX design is typically associated with design elements related to visuals and navigation in digital platforms, TTX includes the full test-taker experience from item design to test administration to score reporting processes. A positive TTX may promote trust between the test-taker and the *Duolingo English Test*. See Ranalli (2021)'s findings that suggest that language learners' experience with technology may influence their trust in technology. The Test Security Framework interacts with the three core assessment frameworks to ensure (a) item security, (b) secure item delivery, test-taker integrity, and data collection, and (c) secure data storage and data privacy. The *Duolingo English Test* ecosystem framework processes, and interactions with TTX and the digitally-informed chain of inferences influence the expected impact associated with social impact and test validity.

2.1 Ecosystem Framework Components

This section discusses the set of ecosystem frameworks illustrated in Figure 1: (1) the Language Assessment (LA) Design Framework, (2) the Expanded Evidence-Centered Design (e-ECD) Framework, (3) the Computational Psychometrics Framework, and (4) the Test Security Framework. TTX factors interact with ecosystem components to ensure that the test lowers barriers to access and promotes a delightful test-taker experience. The Test Security Framework interacts with the LA Design, the Expanded Evidence-Centered Design (e-ECD) and the Computational Psychometrics Frameworks. The LA Design and e-ECD Frameworks interact with the Computational Psychometrics Framework.

2.2 The Language Assessment Design Framework

The LA Design framework includes five key components. First, construct definition (a) considers constructs relevant to university English language proficiency assessment in terms of independent and integrated language skills required for academic, social, and transactional communication (Biber, 2006); (b) leverages the sociocognitive framework to identify valid constructs that are relevant for English language proficiency assessment; and (c) identifies CEFR levels and descriptors (The Council of Europe, 2001, 2020) that inform and are aligned with test items. Second, test item design (a) operationalizes the construct of university English proficiency and (b) uses the CEFR proficiency levels and descriptors and the sociocognitive framework (Mislevy, 2018; Weir, 2005; White et al., 2015) to design construct-relevant item types. *Duolingo English Test* user experience design practices create a delightful test-taker experience, and accessibility and accommodations requirements are considered to support test takers; item(pre-)piloting supports the creation of high quality item types. Third, automated item generation and scoring (a) leverages state-of-the-art, accurate AI; and (b) considers fairness to mitigate issues, such as inappropriate item content and algorithmic bias caused by the AI. Fourth, the evidence-specification activity considers the data available for collection, including (a) construct-relevant data types as proficiency evidence and (b) data pipeline specifications. Fifth, test-taker readiness materials and practice tests contribute to the overall test-taker experience.

2.2.1 Construct Definition

The *Duolingo English Test* is designed to measure English language proficiency (Settles et al., 2020) for the purpose of informing stakeholder admissions decisions at English-medium institutions. Items on the test are designed to measure core independent English language constructs—specifically, Speaking, Writing, Reading, and Listening, and integrated language skills. In terms of types of communication assessed, the *Duolingo English Test* has many test item types (both independent and integrated language skills) that assess university English, including academic, social, and transactional communication. Independent language skills are evaluated in relative isolation (such as, requiring test takers to prepare an essay in writing in response to a short essay prompt). However, advanced English language proficiency as is needed in university settings requires proficiency in integrated language skills. For example, in online course discussion forums, students need to read peer discussion and respond in understandable written form in order to effectively participate in a discussion. In addition, pragmatics plays a key role in appropriate language use in different listening and

speaking contexts (Crystal, 1997; Kasper & Rose, 2002), such as using the appropriate language register for communicating with instructors versus peers. As well, interactional competence (Canale & Swain, 1980; Galaczi & Taylor, 2018) is critical for communicative interaction in different domains and in varying contexts (Bachman & Palmer, 1996; Chalhoub-Deville, 2003). For instance, the interaction of responding in writing to an email from an instructor requires different pragmatic language skills than participating in a course discussion forum with peers. Both interactions are likely to occur in an academic setting. To understand a test-taker's English language proficiency in each context, both interaction types must be assessed. As the *Duolingo English Test* assessment researchers and designers create new item types, constructs being measured are clearly defined. To do this, the *Duolingo English Test* leverages the CEFR (The Council of Europe, 2001, 2020) levels and descriptors, and the sociocognitive framework (Table 1).

The CEFR (The Council of Europe, 2001, 2020) offers a framework that specifies proficiency levels and skill descriptors that can be used to assess English language proficiency. The *Duolingo English Test* item design process is informed by, and aligned with, CEFR levels[‡] (i.e., A1, A2 (Basic User); B1, B2 (Independent User); C1,C2 (Proficient User)) and the qualitative skill descriptors[§] associated with Speaking, Writing, Reading, and Listening (SWRL), Interactional, and Pragmatic domains. The Council of Europe (2001) CEFR asserts that Proficient (C1-level) use of language indicates that a language learner: "can use language flexibly and effectively for social, academic and professional purposes." The implication is that for a test-taker to achieve proficiency at the higher end of the CEFR scale, they need to manage interpersonal situations (such as peer group collaboration). As well, test takers need pragmatic skills (such as, politeness strategies) to use language appropriately across social, academic, and professional contexts. The Council of Europe (2020) introduced an updated, more comprehensive set of modern, interactional and online communication skills required to develop academic English language proficiency. Interactions in academic contexts are likely to be situated in digital environments, such as email, social media, and collaboration networks (e.g., Google Docs, WhatsApp, etc.). To be successful in university contexts, test-takers need to have the linguistic, interpersonal, and pragmatic skill proficiencies to effectively participate in these types of interactions. The *Duolingo English Test* design process considers these factors as designers conceptualize new, innovative independent and integrated item types to assess English language proficiency.

The *Duolingo English Test* also leverages the sociocognitive framework to inform the construct definition as new test item types are developed. The framework asserts that measurement of a domain proficiency (such as SWRL) may be influenced by other skills (such as critical thinking), content knowledge, and intrapersonal (e.g., confidence), neurological (Mislevy, 2018; White et al., 2015), and experiential (Weir, 2005) factors.

[‡] https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale

[§] https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=090000168045bb52

Table 1. Construct definition table illustrating the Duolingo English Test Dictation Task criteria.

Language Constructs, Sociocognitive skills, & Influential performance											Test item activities
Primary construct			Secondary construct, skills, & knowledge					Other influential factors			
Construct	Macro-skill	Micro-skill	SWRL	PGX	INTXL	CT	CK	INTPSL	EXP	NEU	
Listening	Listening comprehension	Pronunciation, vocabulary, & syntactic knowledge	W	✓	NA	✓	✓	✓	✓	✓	A test taker listens to an utterance and types what they hear

Note. SWRL = Speaking, Writing, Reading and Listening English language proficiency domains
PGX = Pragmatic knowledge
INTXL = Interactional knowledge
CT = Critical thinking
CK = Content knowledge
INTPSL = Intrapersonal factors
EXP = Experiential factors
NEU = Neurological factors
NA = not applicable
✓ = potentially applicable

Table 1 illustrates how task criteria—i.e., language constructs and sociocognitive factors—are used to define the construct for *Duolingo English Test* item types. These task criteria are especially important for *Duolingo English Test* innovation. The primary constructs refer to the target independent or integrated language skills (constructs). Secondary constructs, skills, and knowledge are also essential to consider as these may interact with the primary target construct(s)and affect test score interpretation. Other influential factors always "tag along," meaning that these factors may play a role in test-taker performance, independent of the primary construct being assessed. These include intrapersonal (e.g., test-taker confidence), neurological (e.g., executive function), and experiential (e.g., item format familiarity) factors. Intrapersonal, neurological, and experiential factors may be supported through increased test accessibility and accommodations as part of the user experience (UX). Experiential factors may be managed through test readiness resources.

Table 1 illustrates item task criteria using an example of the *Duolingo English Test*'s Dictation Task—a task that requires the test taker to listen to an utterance and write (type) the statement that they heard. Here, the primary target construct assesses listening. However, the test taker must also write down what they heard. This task taps into vocabulary knowledge (such as, understanding vocabulary words and knowing how to spell them) and syntactic knowledge (such as, understanding how vocabulary fits into a larger syntactic structure). Further, it is possible that pronunciation (i.e., understanding the spoken dialect) and pragmatics (i.e., appropriate vocabulary usage) as well as critical thinking and content knowledge may factor into the test taker's ability to process and accurately write down what they have heard. As mentioned earlier, other intrapersonal, experiential and neurological factors always "tag along." While the task is an independent task, *Duolingo English Test* designers are aware that additional facets of the test item may interact with, and influence the test taker's performance on the task.

Table 1 demonstrates, more generally, the process of how *Duolingo English Test* designers define task type constructs. The primary target construct(s) and subconstruct(s) are selected for a new test item. A test item has at least one primary target construct for an independent task for which data (evidence) will be collected from test-taker responses. In addition, secondary (sub)constructs, and additional skills (such as, critical thinking), content knowledge, and other factors (such as intrapersonal factors, e.g., test-taker confidence) may influence test takers' ability to successfully complete a test item. In Table 1, listening (L) is an independent skill and the primary target construct. Identifying secondary facets is important as it can inform the data collection required to assess a skill. The test taker listens to a statement and writes (types) what they heard. Therefore, the secondary construct is writing. Specifically, evidence of listening comprehension is collected in the form of a written response. Data from secondary facets may not always be collected. However, even when it is not explicitly collected, knowledge of secondary facets can support test score interpretation. In addition to informing the construct definition for a test item, Table 1 can support the evidence-specification activity for subsequent data collection and modeling in the e-ECD and Computational Psychometric Frameworks. The table illustrates the relevant data types for which evidence could be collected and used to model test-taker English language proficiency.

2.2.2 Test Item Design For test item design, *Duolingo English Test* item designers operationalize the construct definition as they refresh the content for existing item types and create new item types. As discussed above, to create *Duolingo English Test* item types, designers consider CEFR levels and descriptors and the sociocognitive framework factors associated with university English language proficiency. Primary constructs associated with university English language proficiency for current test item types are illustrated in Table A2. The *Duolingo English Test* continues to innovate and operationalize new and more complex, digitally-mediated item types. As this happens, test designers consider the assessment of a wider set of constructs for independent and integrated language skills relevant to university English. Further, designers consider how to assess different skills in authentic, digitally-mediated settings (for example, online discussion forums or chatbot interactions). As the test continues to innovate and items become more complex and integrated with more varied digitally-mediated facets, test designers consider the full range of sociocognitive framework constructs that may affect test performance (See Table 1). The primary and secondary constructs and other influential factors inform test-taker data (evidence) collection used in the e-ECD and Computational Psychometrics frameworks.

User experience design is essential for item development in order to create delightful experiences for users, as well as to ensure that the test provides accessibility and accommodations for individuals with disabilities. Duolingo is a leader in user experience design. Innovative item types on digital-first assessments are likely to incorporate complex interactions using multiple modalities. *Duolingo English Test* designers who support digital-first assessments continuously iterate on design guidelines to drive design decisions for complex items that incorporate interactions. This process is critical to the *Duolingo English Test* to ensure that the test is generally accessible to everyone. This is essential to the test's support of inclusion and an overall positive TTX. *Duolingo English Test* designers follow federal and industry standards to guide item accessibility and accommodations. Designers continually iterate on item design to advance accessibility and accommodations.

Item piloting is critical to evaluating viability of new item types. Specifically, item piloting informs task design, scoring, and validity early in the development process, and supports longer term innovative item research for *Duolingo English Test*. The test leverages its new pre-pilot platform. For experimental item types, this innovative item pre-piloting method collects item response data from potential test takers. This platform provides a scalable, long-term solution and supports continuous development of new and complex item types. Before new items are added to a section on the test, they undergo a formalized fairness and bias review process using *Duolingo English Test* guidelines. The guidelines leverage and build upon Zieky (2015). Assessment researchers and designers continue to investigate current thinking associated with fairness (such as Randall, 2021). Differential item functioning evaluation is also conducted.

2.2.3 Item Generation and Scoring The *Duolingo English Test* automatically generates test items and automatically scores item responses. Therefore, when a new item type is being conceptualized, the availability of accurate and ethical AI capabilities is a key consideration. There is a significant body of research on automated item generation (Heilman & Smith, 2010; Madnani, Burstein, et al., 2016; Mitkov & Ha, 2003). To our knowledge, the *Duolingo*

English Test is the first large-scale, high-stakes English language assessment that leverages AI to automatically generate test items. In addition to cost savings and efficiencies gained from automated item generation, there are other advantages. For instance, AI can automatically generate new items on a more regular basis than is possible with human item developers. This mitigates test security issues associated with item exposure. Specifically, the ability to continuously generate new item types makes it less likely that different test takers will see the same item on a test. Continued advances in AI such as GPT-3 (Brown et al., 2020) increase the ability to generate items automatically. Automated scoring has been widely used for some time for assessment of constructed-response writing items (Attali & Burstein, 2006; Burstein et al., 1998; Foltz et al., 1998; Madnani, Cahill, et al., 2016; Shermis & Burstein, 2013). The *Duolingo English Test* uses automated scoring for all selected-response and constructed-response scored item types.

It should be noted that while the *Duolingo English Test* does leverage automated item generation and automated scoring, it uses "human-in-the-loop AI." Specifically, as assessment researchers and designers develop new item types, they consider the extent to which the capabilities can support item generation and scoring and what type of human intervention is required. For item generation, for example, human review is used to evaluate fairness and potential bias in automatically-generated test passages and items. For scoring, quality control measures are implemented to detect automated scoring anomalies at scale with the Analytics for Quality Assurance in Assessment (AQuAA) system (see Liao et al., 2021 for details).

2.2.4 Evidence-Specification

The evidence-specification activity supports assessment validity. As new item types are designed, deliberate decisions are made about data collection from test-taker responses. Task criteria, such as those illustrated in Table 1, inform the data collection—specifically, product or process data. Product data are those data derived from the test-taker product (such as, essay writing samples), and process data represent a test-taker's process (such as, test-taker item response duration). This deliberate focus on the test-taker data collected as evidence is aligned with ECD (Mislevy et al., 2003), e-ECD (Arieli-Attali et al., 2019) and computational psychometric principles (von Davier, 2017). This activity ensures that appropriate data (evidence) are collected for subsequent modeling and scoring processes. The evidence selected for data collection and subsequent modeling contributes to the digitally-informed chain of inferences, and ultimately supports the test score interpretation and use. Attention is given to data collection to avoid potential bias (Wise et al., 2021).

The evidence-specification process supports development of a data pipeline, where test-taker data are securely managed—i.e., extracted and stored. The data pipeline interacts with the assessment module in the e-ECD framework and the Computational Psychometrics Framework where raw data are converted into more refined feature measures for test-taker modeling.

2.2.5 Test-taker Readiness Materials & Practice Tests

To support TTX, the *Duolingo English Test* offers test takers a free test readiness guide and practice tests. The readiness guide and practice materials provide important information about types of test tasks, response format, scoring and sample performances allowing potential test-takers to develop experience and to build confidence to take the test. Several brief videos provide different overviews of different

aspects of the test, from the steps to install the *Duolingo English Test* desktop app and how to take the test. The test's website also offers test-takers informational material, such as a list of institutions that accept the test, receipt of the score report, and how to send test results.

Free test preparation include (1) an extensive readiness guide that provides detailed information that expands on content in the short videos and provides some practice materials, (2) a 15-minute practice test that provides an estimated *Duolingo English Test* score and can be taken multiple times, and (3) additional Duolingo online resources in partnership with other organizations, such as World Education Services, and Penn English Language Programs at the University of Pennsylvania offers a free, extensive 11-part *Duolingo English Test* video series. Test-taker preparation and practice tests support a positive TTX. Test prep opportunities for the *Duolingo English Test* exist and are growing. There are free YouTube test prep creators which offer varied information about test content, such as Teacher Luke, EZApply International, Raman offers courses in Hindi, Yulia, and Bruno. Teacher Sally has substantive practice for different parts of the *Duolingo English Test*. Further, Duolingo users self-organize WeChat mini programs for language practice which can support English language development and, in theory, test performance.

2.3 Expanded Evidence-Centered Design Framework

Mislevy et al. (2003)'s evidence-centered design (ECD) is a conceptual framework for building educational assessments that supports an evidentiary argument about student knowledge, skills, and abilities. The framework includes task design, test configuration, development of feature measures, and statistical modeling of relevant measures to generate a student model. The student model supports inferences about student knowledge, skills, and abilities relevant to the assessment. This framework provides tools for planning and creating evidence-centered assessments.

The Expanded Evidence-Centered Design (e-ECD) framework builds on Mislevy's ECD framework. In contrast to Mislevy's ECD framework, the e-ECD framework adds a learning branch that supports formative assessment and instruction (Arieli-Attali et al., 2019) (Figure 2). The *Duolingo English Test* intentionally uses the e-ECD framework because it contains the learning branch. This branch serves as a "placeholder" for longer-term test innovation. This branch provides an opportunity for growth for formative assessment and learning. Currently, the *Duolingo English Test* leverages only the assessment branch. A brief description of the e-ECD framework Task model, the Observational-evidence model and the KSA-model, and how these are applied to the *Duolingo English Test* are provided in this section. The e-Assembly model, in a nutshell, determines how the Task model, Observational-evidence and KSA models work together. Figure 2 illustrates the e-ECD framework internal components and processes, and how the framework interacts with the LA Design Framework.

A Task model contains the test item configuration. Specifically, it contains information about which items will be administered on the test. Decisions about which test items will be administered will determine the set of primary target constructs to be measured, and which test-taker data are collected to support construct measurement (per the LA Design Framework

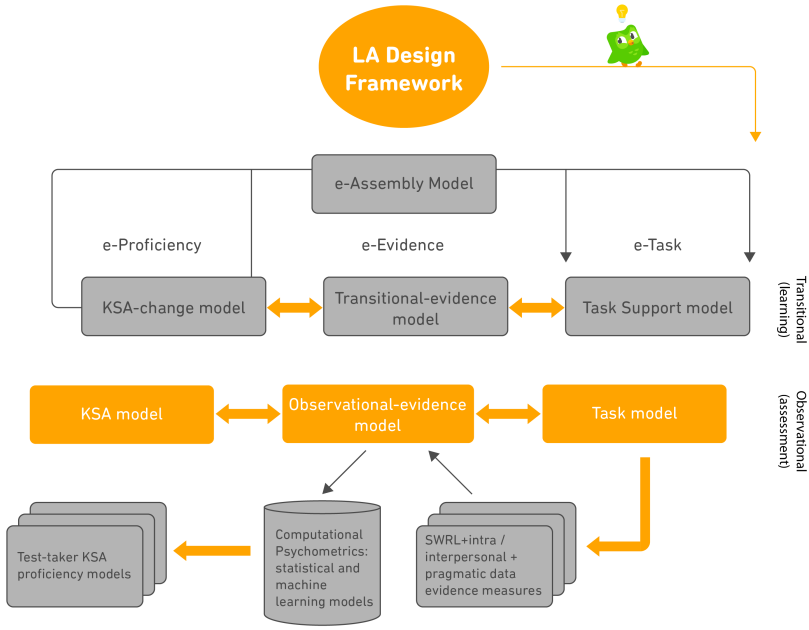

Figure 2. Expanded Evidence-Center Design Framework "interacting" with the LA Design Framework.

specifications). Subsequently, these data are used in the Observational-evidence and KSA models. It is possible to have multiple task models on an assessment if KSAs are related to different constructs. Further, integrated task types that measure multiple constructs might also have their own Task Model.

The Observational-evidence model leverages Computational Psychometrics (discussed in the following section) to create construct-relevant feature measures and to model test-taker proficiency. First, the relevant raw data types from the evidence-specification activity are extracted from the data pipeline (per the LA Design Framework specifications). Data may contain process data (such as, timestamps and keystroke logs) and product data (such as, multiple choice response, and open-ended responses). Through the Computational Psychometric Framework, appropriate statistical and machine learning methods are identified to convert raw data into observable feature measures, such as aggregate features with continuous values. Next, statistical or machine learning methods are applied to the feature measures to generate a KSA model (Knowledge, Skills, and Abilities (KSA)-model).

The KSA-model supports inferences about the test-taker KSAs. Specifically, these are the target constructs measured through the test items. In the KSA model state, a test (informed by the Task Model) has been administered to the test taker. Through completion of the test, appropriate data (evidence) have been collected and modeled (via the Observational-evidence model, and leveraging Computational Psychometrics Framework feature measure and modeling decisions).

The KSA model contains model(s) of test-taker English language proficiency and the test scores from which inferences about test-taker proficiency can be drawn. Test scores are intended to be used by stakeholders to inform their admissions decisions. KSA model information can also be used for validity studies, such as examining relationships between scores and external criteria (such as course grades, or other assessment scores). These types of studies can strengthen test validity.

2.4 Computational Psychometrics Framework

The Computational Psychometrics Framework defines how raw data (evidence) is configured into feature measures, and which measurement procedures should be used to model the evidence to inform test-taker proficiency. It represents an interdisciplinary field that supports the use of AI and machine learning within new psychometric applications, where the data are bigger, richer, and more diverse than in traditional applications.The framework's algorithms and psychometric models are combined to support the test's validity, reliability, and generalizability. In the ecosystem, the framework interacts with the e-ECD, Observational-evidence model. The computational psychometrics framework guides decisions related to feature measures, and statistical and machine learning modeling in the assessment. In this framework, psychometric models can be estimated using the tools developed in computer science for the analysis of many different types of data, including multimodal data, in order to establish how information and evidence can be derived from the data and connected to higher order constructs from the psychometric models.

Computer-based testing collects process data (e.g., time stamp, click stream data) which is all collected indiscriminately and few consider how to design the experience so that the data collected is useful (see Ercikan & Pellegrino, 2017). Similarly, von Davier (2017) argues that the main feature of computational psychometrics is that the data collection is intentional and by design, hence theory-based. In this way computational psychometrics allows researchers to form links between the higher-level abstract models to the concrete components of the fine-grained data in a top-down manner. The machine learning paradigm, on the other hand, allows one to abstract the concrete components in a bottom-up manner by utilizing algorithms to build predictive models given all available data at hand. Computational here means that in order to successfully analyze multimodal big data, and to form the links from this data to higher order abstract constructs, additional analyses that utilize computational models (from ML to statistical/psychometrics) are required (LaFlair et al., n.d.). AI considerations are significant for this framework since it involves algorithmic and statistical modeling that influence the KSA model (i.e., test-taker proficiency) and the final score use. To that end, it is essential that the full method—i.e., data and modeling methods—is audited to mitigate potential inequities (Wagner et al., 2021). The data modeling methods must be evaluated to ensure that they are fair (e.g., do not disadvantage a subpopulation), transparent and explainable, and there is a human in the loop to review algorithmic decisions (see Liao et al., 2021).

3 Test Security Framework

The Test Security Framework is responsible for all aspects of ensuring that the test is securely delivered and scored to mitigate situations such as test imposters, leaked test items and test-taker integrity. The test security framework spans across the three key assessment frameworks (See Figure 1). The test uses a desktop application using the Electron framework—a framework for developing cross-platform desktop apps. By having an application installed on the test taker's computer, it is possible to flag potential test-taker integrity issues by getting operating system–level signals. These are strong signals that indicate potential compromises in test-taker integrity via changes to the test-taker's environment, including recognizing suspicious software running on their computer at the time of the test and connected peripherals. A team of human proctors maintains 24/7 proctoring coverage to meet the needs of its global user base. The proctoring team is trained in ID verification and detecting/spotting suspicious behaviors, while also ensuring fairness and functionality of the test itself. In addition, human proctors participate in regular calibration activities, monthly meetings, discussion groups, and cultural bias training to minimize bias and ensure fairness in proctoring. Escalations and edge cases are handled by more experienced senior proctors. When a unique test situation presents itself, a manager makes the final decision on the test certification. The *Duolingo English Test* proctoring team also strives for efficiency to ensure that test-takers receive their exam results within 48 hours.

Test security interacts with TTX with regard to design, evidence capture, and proficiency modeling and score reporting. With regard to design, there are considerations such as test-taker experience and accessibility. For instance, does the test security environment cause anxiety, or is the set up seamless and unintimidating? Does the test security in any way interfere with accessibility? In the context of collection of evidence capture, does facial recognition bias prevent test-takers from starting the test, or might the test security infrastructure hiccup and cause loss of response data? Further, are data being collected and stored responsibly so as not to compromise data security and privacy issues? From a proficiency modeling perspective, does test security in any way compromise fairness (e.g., data loss)? For score reporting, is the proctoring process accurate in mitigating test-taker integrity issues? Is proctoring efficient and accurate, so that test-takers can be assured the quick-turnaround time for test results? Considerations, such as those suggested above, are critical to achieve expected impact associated with TTX and test score validity.

4 Discussion

This paper presents the *Duolingo English Test*'s novel theoretical assessment ecosystem. In contrast to the previous frameworks, the *Duolingo English Test* ecosystem is composed of a coherent, comprehensive, and integrated set of complex assessment frameworks that addresses the prominent digital attributes of the test. The ecosystem frameworks include: (1) the Language Assessment Design Framework, (2) the Expanded Evidence-Centered Design (e-ECD) Framework, (3) the Computational Psychometrics Framework, and (4) the Test Security Framework. TTX spans across the entire ecosystem. TTX considerations include factors such as low price point and shorter testing time, free test-readiness resources, delightful UX

design, accessibility and accommodations, and fast score turn-around processes. The paper also introduces a novel, digitally-informed chain of inferences that interacts with the ecosystem. The Duolingo English Test's digitally-informed chain of inferences adapts Chapelle et al. (2008)'s proposed a chain of inferences supporting a validity argument for a high-stakes English language assessment. Further, the Duolingo English Test digitally-informed chain of inferences expands on digital considerations for a chain of inferences proposed in Xi et al. (2008) and Xi (2010), respectively, for automated speech and essay scoring used on assessments. The *Duolingo English Test* ecosystem and the digitally-informed chain of inferences support the test's expected impact associated with Duolingo's social mission and test score validity, and drive *Duolingo English Test* innovation. The test aims to be valid, reliable and fair and consider the impact of technology (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014).

The *Duolingo English Test* ecosystem and digitally-informed chain of inferences is consistent with Van Moere and Downey (2016)'s assertion that assessments need to consider technology and AI to build validity arguments. In a similar spirit, Cope et al. (2020) present contrasting "traditional assessment artifacts and e-learning ecologies where artificial intelligence has exploited new opportunities for the processes of learning" (p. 12). In contrast to AI-enabled assessments, they define traditional assessments as paper-and-pencil and first-generation, computer-based assessments that do not necessarily include AI. They offer examples of traditional and AI-enabled assessments. They characterize traditional assessments as: "Peculiar artifacts at distinct times: select response and supply response tests. Even when they are more frequent (e.g. quizzes after a video lecture or the end of the chapter in an e-textbook), such tests remain summative in their genre and orientation." and AI-enabled assessments as: "Embedded formative assessment: measurement of learning that offers incremental, semantically legible, machine feedback and machine-mediated human feedback." (Cope et al., 2020, p. 13). Cope et al. (2020)'s ideas are aligned with digital-first assessment and food-for-thought for modern, AI-enabled assessment. Especially in the case of digital-first assessment, it is essential to identify new opportunities that are available to support AI-enabled learning and assessment. Further, it is important to consider how those opportunities can be leveraged to build innovative digitally-mediated, construct-relevant test items, and thoughtfully collect evidence that can be effectively modeled to provide meaningful information about test-taker proficiency.

The *Duolingo English Test* ecosystem was designed to represent the key processes for building a secure, valid, fair, and reliable English language proficiency assessment that can be used by stakeholders at English-medium institutions to inform admissions decisions. The ecosystem is flexible. Its components can be modified and expanded to accommodate on-going test developer insights, and test-taker and test user needs. Further, digital considerations associated with the chain of inferences can be updated concurrent with innovation. Innovative considerations, for instance, might expand how the *Duolingo English Test* addresses fairness and potential bias in current socio-political contexts (such as Randall, 2021), or how advances in AI are deployed on the test. The ecosystem is intended to drive continuous *Duolingo English Test* innovation, while also maintaining Duolingo's social mission and facilitating expected impact.

Acknowledgements

We would like to thank our Duolingo colleagues, Will Belzak, Ramsey Caldwell, Sarah Goodwin, Jeff Tousignant and Sophie Wodzak for discussion, and insightful comments on earlier versions of this paper. Additional thanks go to Sophie Wodzak for copyediting. Thanks to Basim Baig and Jeff Tousignant for providing content, respectively, about test security and test readiness resources. We thank Micheline Chalhoub-Deville who provided invaluable on-going discussions and insights as we prepared this paper, and Steve Sireci for bringing to our attention a modern, antiracist framework of test item development.

5 References

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for educational & psychological testing*. American Educational Research Association, American Psychological Association & National Council on Measurement in Education.

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.

Association for Test Publishers. (2021). *Artificial intelligence and the testing industry: A primer, a special publication from ATP: Association for test publishers*.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, *4*(3).

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Barrett, M. D., Jiang, B., & Feagler, B. E. (2021). A smart authoring system for designing, configuring, and deploying adaptive assessments at scale. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-021-00258-y.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. John Benjamins.

Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America*, *112*(1), 272–284.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Amodei, D. (2020). *Language models are few-shot learners*.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998). Automated scoring using a hybrid feature identification technique. *Proceedings of the Annual Meeting of the Association of Computational Linguistics*.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, *20*, 369–383.

Chalhoub-Deville, M. (2009). The intersection of test impact, validation, and educational reform policy. *Annual Review of Applied Linguistics*, *29*, 118–131.

Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity: Theoretical development and integrated arguments*. Equinox Publishing Limited.

Chapelle, C., Enright, M., & Jamieson, J. (2008). *Building a validity argument for the test of english as a foreign language*. Routledge.

Cope, B., Kalantzis, M., & Searsmith, D. (2020). Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. *Educational Philosophy and Theory*, 1–17.

Crystal, D. (Ed.). (1997). *The cambridge encyclopedia of language*. Cambridge University Press.

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

Ercikan, K., & Pellegrino, J. W. (Eds.). (2017). *Validation of score meaning for the next generation of assessments: The use of response processes*. Taylor & Francis.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI*. Berkman Klein Center for Internet & Society. http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *25*, 285–307.

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, *15*(3), 219–236.

Heilman, M., & Smith, N. A. (2010). Good question! Statistical ranking for question generation. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 609–617. https://aclanthology.org/N10-1086

Jessop, L., Suzuki, W., & Tomita, Y. (2007). Elicited imitation in second language acquisition research. *Canadian Modern Language Review*, *64*(1), 215–238.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527.

Kane, M. T. (2011). Book review: Language assessment in practice: Developing language assessments and justifying their use in the real world. *Language Testing*, *28*(4), 581–587. https://doi.org/10.1177/0265532211400870

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kasper, G., & Rose, K. R. (2002). *Pragmatic development in a second language*. Basil Blackwell.

Khodadady, E. (2014). Construct validity of c-tests: A factorial approach. *Journal of Language Teaching and Research*, *5*(6), 1353.

Klein-Braley, C. (1997). C-tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, *14*(1), 47–84.

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

LaFlair, G. T., Yancey, K. P., Settles, B., & von Davier, A. A. (n.d.). *Computational psychometrics for digital-first assessments: A blend of ML and psychometrics for item generation and scoring* (Y. V. & M. von Davier, Eds.).

Liao, M., Attali, Y., & von Davier, A. A. (2021). AQuAA: Analytics for quality assurance in assessment. *Proceedings of the Educational Data Mining Conference (Virtual*, 787–792.

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.

Madnani, N., Burstein, J., Sabatini, J., Biggers, K., & Andreyev, S. (2016). Language MuseTM: Automated linguistic activity generation for english language learners. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Madnani, N., Cahill, A., & Riordan, B. (2016). *Automatically Scoring Tests of Proficiency in Music Instruction in Proceedings of the Eleventh Workshop on Innovative Use of NLP for Building Educational Applications*, 217–222,.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–105). Macmillan.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5–8.

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. *Communicative Proficiency and Linguistic Development: Intersections Between SLA and Language Testing Research*, 211–232.

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in english as a foreign language. In *Insights into non-native vocabulary teaching and learning* (pp. 83–98). Multilingual Matters.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), –29.

Mitkov, R., & Ha, L. A. (2003). Computer-aided generation of multiple-choice tests. *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, 17–22.

Papageorgiou, S., Davis, L., Norris, J. M., Gomez, P. G., Manna, V. F., & Monfils, L. (2021). *Design framework for the TOEFL essentialsTM test*.

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing*, *52*, 100816.

Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*, *0*(0), 1–9.

Reichert, M., Keller, U., & Martin, R. (2010). *The c-test, the TCF and the CEFR: A validation study. Der c-test: Beiträge aus der aktuellen forschung* (pp. 205–231). Contributions from Current Research.

Settles, B., LaFlair, G. T., & Hagiwara, M. (2020). Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, *8*, 247–263.

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and future directions*. Routledge.

Staehr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*(2), 139–152.

Tannenbaum, R. J., & Katz, I. R. (2021). *Validity considerations in complex task design* (Vol. 5, pp. 196–214). The Journal of Writing Analytics. https://doi.org/10.37514/JWA-J.2021.5.1.06

The Council of Europe. (2001). *Common european framework of reference for languages: Learning, teaching, assessment*. Press Syndicate of the University of Cambridge.

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.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, *29*(3), 325–344.

Van Moere, A., & Downey, R. (2016). In *Technology and artificial intelligence in language assessment. 10.1515/9781614513827-023*.

Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, *12*(1), 54–73.

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

von Davier, A. A. (2021). Research-based digital-first assessments and the future of education. *Proceedings of Artificial Intelligence in Education: 22nd International Conference, (AIED)*.

Wagner, C., Strohmaier, M., Olteanu, A., Kıcıman, E., Contractor, N., & Eliassi-Rad, T. (2021). Measuring algorithmically infused societies. *Nature*, *595*, 197–204. https://doi.org/10.1038/s41586-021-03666-1

Weir, C. J. (2005). Language testing and validation. *Hampshire: Palgrave McMillan*, *10*, 9780230514577.

White, E. M., Elliot, N., & Peckham, I. (2015). *Very like a whale: The assessment of writing programs*. University Press of Colorado.

Wise, A. F., Sarmiento, J. P., & Boothe Jr, M. (2021). Subversive learning analytics. *Lak21: 11th International Learning Analytics and Knowledge Conference*, 639–645.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, *27*(3), 291–300.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). Automated scoring of spontaneous speech using SpeechRaterSM v1. 0. *ETS Research Report Series*, *2008*(2), –102.

Zieky, M. J. (2015). Developing fair tests. In *Handbook of test development* (pp. 97–115). Routledge.

Zimmerman, J., Broder, P. K., Shaughnessy, J. J., & Underwood, B. J. (1977). A recognition test of vocabulary using signal-detection measures, and some correlates of word and nonword recognition. *Intelligence*, *1*(1), 5–31.

A Appendix

Table A1. Digitally-informed chain of inferences

Inference	Warrant	Example Underlying Theoretical Assumptions & Associated Digital Considerations			
		Design	e-ECD	Computational Psychometrics	Test Security
Domain description	The *Duolingo English Test* item types represent knowledge, skills and abilities associated with constructs relevant to university English language skills required for English-medium institutions, including digitally-mediated communication	Assumption: English language skills require digitally-mediated interactions in academic settings. Digital consideration: Does digital item design reflect an authentic, construct-relevant interaction?	Assumption: English language skills can be accurately identified. Digital consideration: Are sufficiently accurate statistical / ML methods available to accurately identify construct-relevant English language skills associated with digital task interaction?	Assumption: English language skills can be accurately identified with regard to bias. Digital consideration: Are statistical / ML methods available to ethically identify construct-relevant English language skills?	Not applicable

(continues)

Table A1. (continued)

Inference	Warrant	Example Underlying Theoretical Assumptions & Associated Digital Considerations			
		Design	e-ECD	Computational Psychometrics	Test Security
Scoring	Observed *Duolingo English Test* performance based on automated evaluation methods is reflective of university English language skills required for English-medium institutions.	Assumption: Digitally-mediated Item response product and process data reflect English language skills. Digital consideration: Are AI methods available that can generate construct-relevant feature data?	Assumption: English language skills measures can be extracted from raw digitally-mediated Item response product and process data. Digital consideration: Are AI methods that generate raw feature data sufficiently accurate to produce relevant and accurate feature measures?	Assumption: English language skills measures developed from raw digitally-mediated Item response product and process data. Digital consideration: Are statistical/machine learning methods sufficiently accurate to generate feature measures representing English language skills?	Assumption: Test-taker identity is accurately identified. Digital consideration: Are digital test security measures sufficiently accurate to identify the test taker?

(continues)

Table A1. (continued)

Inference	Warrant	Example Underlying Theoretical Assumptions & Associated Digital Considerations			
		Design	e-ECD	Computational Psychometrics	Test Security
Generalization	Observed *Duolingo English Test* performance measures are estimates of expected performance for parallel versions of an automatically-generated test, and across automated and human raters and test administrations.	Assumption: Parallel versions of item types can be designed to assess English language skills requiring digitally-mediated interactions in academic settings. Digital consideration: For parallel versions of item types requiring digitally-mediated interactions, does data capture produce consistent measures, especially if data types are varied?	Assumption: Test scores from simulated & digitally-mediated interactions can be measured reliably. Digital consideration: Will AI methods applied to extract features across varying digitally-mediated interactions produce reliable feature measures?	Assumption: Test scores from simulated & digitally-mediated interactions can be measured reliably. Digital consideration: Will statistical/machine learning methods applied to model test-taker KSAs across varying digitally-mediated interactions produce reliable models?	Assumption:Test scores from simulated & digitally-mediated interactions can be measured reliably. Digital consideration: Are digital test security measures sufficiently accurate to identify the test taker in a test-retest situation?

(continues)

Table A1. (continued)

Inference	Warrant	Example Underlying Theoretical Assumptions & Associated Digital Considerations			
		Design	e-ECD	Computational Psychometrics	Test Security
Transparency & Explanation	Observed *Duolingo English Test* performance provides interpretable English language proficiency measures consistent with university English language skills required for English-medium institutions.	Assumption: Data collected to produce performance scores has a clear mapping to construct-relevant task criteria. Digital consideration: Are AI methods available that can generate construct-relevant, explainable, feature data?	Assumption: Data collected to produce performance scores has a clear mapping to construct-relevant task criteria. Digital consideration: Do AI methods accurately produce construct-relevant, feature data that can be objectively evaluated with statistical measures?	Assumption: Data collected to produce performance scores has a clear mapping to construct-relevant task criteria. Digital consideration: Do statistical/machine learning models contain traceable, construct-relevant, feature measures that support a clear explanation of test-taker performance?	Not applicable

(continues)

Table A1. (continued)

Inference	Warrant	Example Underlying Theoretical Assumptions & Associated Digital Considerations			
		Design	e-ECD	Computational Psychometrics	Test Security
Extrapolation	The *Duolingo English Test* assesses the construct of English language proficiency consistent with university English language skills required for English-medium language institutions.	Assumption: Data collected to produce performance scores accurately represent construct-relevant task criteria. Digital consideration: Are AI methods available that can generate construct-relevant feature data (evidence) that can be examined in relation to external measures?	Assumption: Data collected to produce performance scores are related to relevant external measures of academic proficiency. Digital consideration: Do AI methods accurately produce construct-relevant, feature data (evidence) that can be objectively evaluated in relation to external measures?	Assumption: Data collected to produce performance scores are related to relevant external measures of academic proficiency. Digital consideration: Do statistical/machine learning models of test-taker performance produce sufficiently accurate measures so they can be evaluated in relation to external measures?	Not applicable

(continues)

Table A1. (continued)

Inference	Warrant	Example Underlying Theoretical Assumptions & Associated Digital Considerations			
		Design	e-ECD	Computational Psychometrics	Test Security
Use of test scores	Observed *Duolingo English Test* performance is beneficial for stakeholders.	Assumption: Test users can leverage test score reports to support key decisions.	Assumption: Test users benefit from valid test scores that support key decisions.	Assumption: Test users benefit from valid test scores that support key decisions.	Assumption: Test users benefit from valid test scores that support key decisions.
		Digital consideration: Can stakeholders meaningfully interpret digitally-generated features associated with test scores—i.e., interpret test score reporting measures?	Digital consideration: Do AI methods accurately produce construct-relevant, feature data rendering usable test scores?	Digital consideration: Are statistical/machine learning methods models sufficiently ethical and accurate rendering usable test scores?	Digital consideration: Are digital test security measures sufficiently accurate so that there is stakeholder confidence about test-taker identity, and test score use confidence?

Table A2. Item types on the Duolingo English Test

Item name	Activity	Primary Target Construct(s)	Integrated Skills (LaFlair, 2020)	Type of scoring	Average number of items	References
Vocabulary Yes/No	Read and select English words	R,W	Literacy Comprehension	CAT	6	Milton (2010); Staehr (2008); Zimmerman, Broder, Shaughnessy, & Underwood (1977)
Vocabulary Yes/No	Listen and select English words	L, S	Conversation Comprehension	CAT	6	Milton et al. (2010); Milton (2010)
C-Test	Read and complete words	R, W	Literacy Comprehension	CAT	6	Khodadady (2014); Klein-Braley (1997); Reichert, Keller, & Martin (2010)
Dictation	Listen and Write	L, W	Conversation Comprehension	CAT	6	Bradlow & Bent (2002 & 2008)
Elicited Imitation	Read aloud	R, S	Conversation Comprehension	CAT	6	Jessop, Suzuki, & Tomita (2007); van Moere (2012); Vinther (2002)
Short writing	Write about the photo	W	Literacy Production	Performance	3	Cushing-Weigle (2002)

(continues)

Table A2. (continued)

Item name	Activity	Primary Target Construct(s)	Integrated Skills (LaFlair, 2020)	Type of scoring	Average number of items	References
Extended writing	Write your response	W	Literacy Production	Performance	1	Cushing-Weigle (2002)
Extended speaking	Speak about the photo	S	Conversation Production	Performance	1	Luoma (2004)
Extended speaking	Read and Speak	S	Conversation Production	Performance	2	Luoma (2004)
Extended speaking	Listen and Speak	S	Conversation Production	Performance	1	Luoma (2004)
Video Interview	Video sample	S		Unscored	1	
Extended Writing	Writing sample	E		Unscored	1	