Analysis of the Validity, Design and Development of the Duolingo English Test

Table of Contents

Section 1: Purpose of this Document

The purpose of this document is to report on our analysis of validity research conducted by the Duolingo English Test (DET). We used the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) and DET's documentation for our analysis of the questions mentioned below. We also considered the interpretation/use argument of Kane (2013).

Evidence from Standards

The *Standards* address validity across various chapters, and, consequently, this document integrates sections from the following chapters (chapters 1, 3, 4, 8, 9, and 12). These chapters describe validity evidence from various perspectives including:

- an overview of the test purpose
- information on score use, score interpretation, and score meaning
- cautionary notes regarding score misuse
- description of the intended populations such as populations' demographic information and the conditions under which the data were collected
- evidence supporting score use to make predictions related to future behaviors

A key consideration is articulated in:

> Standard 1.0: Clear articulation of each intended test score interpretation for a specific use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided (AERA, APA, & NCME, 2014, p. 23).

Questions Analyzed

We analyzed the following questions related to the validity of DET score–based inferences:

1. How are DET test scores intended to be interpreted and used?
    a. Are these uses clearly and coherently presented?
    b. What is the evidence supporting these uses and decisions?

2. For which population(s) is the DET intended?
    a. Are the populations clearly and explicitly described?

3. What constructs does the DET measure?
    a. Are the construct and sub-constructs well defined and operationalized?

   b. Are the construct and sub-constructs defined specifically for the purpose of the test and the context in which the test is used?
   c. Are the constructs linked to theories of language learning and assessment?

4. Does the DET content adequately sample from the intended constructs' universe (or domain) to support score interpretations?
   a. What theoretical or empirical evidence has been gathered to support that the DET's content is representative of the defined construct?

5. Does the internal structure of the DET provide support for score interpretations?

6. What is the relationship of DET scores to other related variables? Does the strength of these relationships provide support for score interpretations?

7. Do consequences stemming from the use of the DET scores in decision-making provide support for score interpretations?

These questions are central to an interpretation/use argument (IAU; Kane, 2013) demonstrating the validity of inferences based on Duolingo English Test scores.

The Interpretation/Use Argument

The IUA specifies the sequence of inferences and assumptions involved in moving from a test taker's performance (score) to the decisions made from scores. Kane (2013) discusses that a proposed test score interpretation and use can be considered valid to the extent that the IUA is coherent and complete and its assumptions are supported by evidence. Kane highlights that test developers and users have the responsibility for evaluating the IUA.

In what follows, we summarize our analysis. We also use examples from DET's documentation to illustrate ways in which the DET addressed these questions in their assessment design choices. The goals are to further strengthen the validity of inferences made from the DET for its intended uses and target populations while remaining attentive to the IUA's objectives.

Back to Table of Contents

Section 2: Sources Reviewed

To address the aforementioned goals and questions, we reviewed publicly-available DET documentation (e.g., articles and DET websites for test takers and institutions). We also reviewed a few independently completed evaluations of the DET and held semi-structured interviews with DET staff. The staff included the chief of assessment and assessment scientists. The published resources used in the development of this report are cited in the reference section at the end of this document.

Back to Table of Contents

Section 3: Introduction – Validation and DET

The growth of digital technology and advances in automated scoring has provided the opportunity to develop innovative forms of technology-based assessments. These assessments are designed to measure constructs such as language proficiency in increasingly novel ways. Such advances present opportunities to provide learners with just-in-time feedback through automated scoring of items and more efficient tests. To take advantage of those innovations in principled ways, Digital-first Assessments (von Davier, 2020) require the analysis of both design-based and interpretation-based aspects of validity.

Standards for Analyzing Validity

"Validity refers to the degree to which evidence and theory support test score interpretation for the proposed uses of tests" (AERA, APA, & NCME, 2014, p. 11) and is the most fundamental consideration in developing and evaluating tests.Furthermore, validity concerns the interpretation(s) and use(s) of test scores, rather than the test itself. During the test validation process, relevant evidence is accumulated to provide a scientific basis for the proposed score interpretations and uses.

The *Standards* (AERA, APA, & NCME, 2014) identify five sources of validity evidence:

1. test content
2. response processes
3. internal structure
4. relationship of scores to other variables
5. consequences of testing.

The *Standards* conclude by detailing that not all sources of evidence are required for every score interpretation or use. Researchers should carefully assemble relevant evidence from appropriate sources to develop a scientifically-based IUA.

DET – a Groundbreaking Test

The DET was designed to achieve multiple goals. The primary goal was to develop a test of English language proficiency that provides test takers with an efficient, convenient, and economical means of demonstrating language proficiency. In addition, the designers wanted the test to have **equivalent psychometric quality** to other language tests used for high-stakes decision making. Psychometric quality was defined as having clear evidence-based score

interpretations supporting uses (e.g., admissions decisions), while also having high levels of score reliability/precision.

The designers defined efficiency, convenience, and economy as

- enabling test takers to test anywhere on their personal computers at any time
- making traveling to a test center unnecessary
- reducing test time to one hour or less
- providing scores to test takers in approximately two days
- pricing the test so that it was within reach of nearly any motivated student (e.g., having the price point well below the cost of alternative tests)

Achieving these five goals required the DET developers to move beyond traditional test development processes (Settles, LaFlair, & Hagiwara, 2020). DET utilized machine learning (ML) and natural language processing (NLP) technology to accomplish their goals. Their goals further required them to adopt a computer-adaptive test (CAT) model. To develop a sufficiently large item pool to ensure test security required them to develop an automated item generation process, which not only developed items, but provided item difficulty estimates. In so doing, the developers decided to eliminate traditional item pretesting and instead relied on the ML and NLP technology to build the needed item pool and attain the desired price point.

Using ML and NLP technology to achieve a large CAT item pool without pretesting **revolutionizes test development**. This groundbreaking action in testing is termed a **Digital-First Assessment** (von Davier, 2020). Digital-first assessments are characterized by

- Computational psychometrics as an integrative framework
- Computational models and data-driven algorithms for test development
- Artificial Intelligence (AI)-based tools
- Comparability of test scores

Validation and the DET

In the following sections, we develop an IUA supporting the use of the DET score interpretations. By providing evidence that answers the questions defined in the “Purpose of This Document” section, an IUA is constructed. In answering these questions, we are continually mindful of the innovative nature of the DET as a Digital-First Assessment. Traditional psychometric models and interpretations may need to be modified to accommodate these innovative breakthroughs. Evaluators may need to analyze the preponderance of the evidence and assess whether it collectively provides sufficient theoretical and empirical basis supporting test score claims and uses.

Back to Table of Contents

Section 4: Intended Uses and Score-Based Interpretations

Question: How are DET test scores intended to be interpreted and used?

a. Are these uses clearly and coherently presented?

b. What is the evidence supporting these uses and decisions?

Standard 1.0: Clear articulation of each intended test score interpretation for a specific use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided (AERA, APA, & NCME, 2014, p. 23).

Standard 1.1: The test developer should set forth clearly how test scores are intended to be interpreted and consequently used. The population for which a test is intended should be delimited clearly, and the construct or constructs that the test is intended to assess should be described clearly (AERA, APA, & NCME, 2014, p. 23).

Within the purpose statement, the *Standards* call for test developers to define whether score interpretations are norm- or criterion-referenced (AERA, APA, & NCME, 2014, p. 76). Norm-referenced score interpretations are designed to compare a test taker's score to scores in a defined population of test takers. Norm-referenced scores allow a test taker to determine if he or she scored above or below the average score of the defined population. Criterion-referenced scores allow individuals to determine whether they have mastered a defined set of knowledge or skills or a set of defined learning standards.

DET Purpose Statement and Score Interpretations

In this section, we summarize our review of DET's documentation (e.g., articles and website) with regards to the uses of the DET (e.g., admissions to undergraduate and graduate institutions and placement and exit decisions) and the information (other tests and data sources) available to DET users to inform score-based decisions and interpretations. In our review, we considered the literature on language proficiency tests and validity and the *Standards*.

The Purpose of the DET. The first paragraph of the DET's Technical Manual provides a summary of its purpose:

The Duolingo English Test is a measure of English proficiency. The test has been designed for maximum accessibility; it is delivered via the internet, without a testing center, and is available 24 hours a day, 365 days a year. It has been designed to be efficient. It takes less than one hour to complete the entire process of taking the test (i.e., onboarding, test administration, uploading). It is a computer-adaptive test (CAT), and it uses item types that provide maximal information about English language proficiency. It is designed to be user-friendly; the onboarding, user interface, and item formats are easy to interact with (LaFlair & Settles, 2020, p. 3).

DET Score Interpretations. Moritoshi (2001) adds that central to the use of language proficiency tests is to evaluate the extent to which the scores provide valuable information to guide test users' focal decisions.

DET scores are used to inform undergraduate- and graduate-level admissions decisions as well as placement and exit decisions from English language programs. DET scores may be used to fulfill institutions' English proficiency admissions requirements. As test scores are aligned to the Common European Framework of Reference (CEFR; we further elaborate on the CEFR in later sections in this report), the DET yields criterion-referenced score interpretations (Brenzel & Settles, 2017; Settles et al., 2020).

The DET is designed to measure English language proficiency related to **speaking, listening, reading,** and **writing** skills (LaFlair & Settles, 2020). To provide a more personalized view of themselves, applicants submit a video interview and writing sample to supplement their test scores on these skills.

The DET also reports test takers' subscores [insert link] and associated CEFR–derived can-do statements. (Appendix A presents the CEFR-derived can-do statements.) Results reports provide learners with information on their use of language skills in reading, listening, speaking, and writing; see the scoring section and associated sub-scores for further details regarding the interpretation of test scores and subscores [insert link].

For additional information related to the intended score interpretations and supporting evidence, also see:

- LaFlair and Settles (2020) describe that the DET is a measure of English language proficiency for communication in English-medium settings

- LaFlair (2020) describes that the DET is used to inform admissions decisions at the undergraduate or graduate level and placement and exit decisions into university English language programs
- Ishikawa, Hall, and Settles (2016), in their two-year study conducted with DePauw University, provide evidence of the use of the DET for informing admissions and placement decisions in academic settings. The study examined the relationship between the DET and the academic English ability of international students. Correlations of DET scores to on-campus faculty assessments of English ability for incoming international students indicated moderate to strong relationships ($r = 0.62^{***}$ for written ability and $r = 0.49^{***}$ for oral comprehensibility). These findings suggest that DET scores are predictive of faculty members' decisions to place students into academic English support classes.

Back to Table of Contents

Section 5: Populations Assessed by the DET

Question: For which population(s) is the DET intended?

a. Are the populations clearly and explicitly described?

The DET is used by different users and stakeholders, including students and institutions to inform decisions regarding admitting students to academic institutions or placing students in courses or language programs. Along this line, the users of the DET are:

- **Undergraduate- and graduate-level applicants** applying to English-medium post-secondary institutions
- **Undergraduate- and graduate-level institutions** reviewing candidates' applications to inform admissions decisions into degree programs taught in English
- **Students** in English language programs
- **Instructors and institutions** in English language programs wishing to:
  - ✔ Monitor students' progress
  - ✔ Inform exit decisions
  - ✔ Identify and validate English language–learning students' level of proficiency with an external test

Test takers provide background information when registering to take the DET during the onboarding process. They provide information on their first language, date of birth, and gender. ID-issuing country (based on the proof of identification) and IP address country are also recorded for each test taker. Collecting this type of information enables DET developers to conduct fairness analyses using the background information to determine the extent to which the test functions similarly for the various test-taker groups [see the following link for additional information on fairness].

Back to Table of Contents

Section 6: Constructs Assessed by the DET

Question: What constructs does the DET measure?

a. **Are the construct and sub-constructs well defined and operationalized?**
b. **Are the construct and sub-constructs defined specifically for the purpose of the test and the context in which the test is used?**
c. **Are the constructs linked to theories of language learning and assessment?**

Haladyna and Downing (2004) note that "The most fundamental step in validation is defining the construct" (p. 17). Along this line, the *Standards* argue that defining test constructs and content specifications necessitate the articulation of the knowledge, skills, abilities, and other attributes the test is designed to measure. Designers delineate the aspects, extent, and limitations of content coverage guided by theory and analysis (AERA, APA, & NCME, 2014). Thus, in what follows we elaborate on DET's construct definition.

DET Construct Definition

The DET is designed to measure integrated English reading, writing, listening, and speaking skills in alignment with the Common European Framework of Reference (CEFR). The CEFR (Council of Europe, 2001, 2018, 2020) is an international standard for describing language ability. It describes language ability on a six-point ordinal scale, from A1 for beginners to C2 for individuals who have achieved high proficiency in a language. It provides a framework for interpreting and comparing scores on various language assessments.
Our goal is to create a test integrating reading, writing, listening, and speaking skills into a single overall score that corresponds to CEFR-derived ability (Settles et al, 2020, p. 249).

Aligning the DET to the CEFR requires that test takers demonstrate their abilities to use English in reading, writing, listening, and speaking within a variety of topics and genres for varying purposes (Settles et al., 2020).

The DET provides a total score along with subscores measuring Literacy, Comprehension, Conversation, and Production (DET website, 2020). These content domains are defined as:

Literacy: the test taker's ability to read and write.
Comprehension: the test taker's ability to read and listen.
Conversation: the test taker's ability to listen and speak.
Production: the test taker's ability to write and speak.

The CEFR provides a common basis for the elaboration of language syllabi, curricular guidelines, examinations, and textbooks across Europe and beyond (Byram & Parmenter, 2012). These abilities it defines are the ability to understand written and spoken language from varying topics, genres, and linguistic complexity and writing or speaking on a variety of topics and for a variety of purposes.

The CEFR uses six Common Reference Levels: A1, A2 (Basic); B1, B2 (Independent); C1, and C2 (Proficient User). "Can-do statements" are used to describe what learners can do. For example, the A2-level contains statements such as "can understand personally relevant phrases and high-frequency vocabulary" (Council of Europe, 2001).

Back to Table of Contents

Section 7: Evidence Based on Test Content

Question: Does the DET content adequately sample from the intended constructs to support score interpretations?

a. What theoretical or empirical evidence has been gathered to support that the DET content is representative of the defined construct?

> Standard 1.11: When the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified with reference to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency or criticality, these criteria should be clearly explained and justified (AERA, APA, & NCME, 2014, p. 26).

Decisions regarding item and response formats must align for the purposes of the test, the test's defined domain, and the testing platform. The selection of item types is frequently determined by considerations of the testing platform and scoring ease; however, validity issues also must remain a central part of the decision process (AERA, APA, & NCME, 2014). Additionally, "designing tests to be accessible and valid for all intended examinees, to the maximum extent possible, is critical" (AERA, APA, & NCME, 2014, p. 77).

DET and the Operationalized Content

DET operationalized the construct through ten different item formats. Collectively, the formats assess reading, writing, listening, and speaking using different forms. The item types neither sample the full domain of language use nor require test takers to demonstrate all the linguistic tasks relevant to the CEFR levels. Instead, they serve as proxies for the underlying skills. The item type selection choices the DET made reflect the following considerations: the item formats can "be automatically generated and graded at scale, and have decades of research demonstrating their ability to predict linguistic competence" (Settles et al., 2020, p. 249). (Appendix B provides citations to studies and papers supporting the use of the item formats.) Table 1.1 provides a list of the ten DET item formats along with the skills measured by the item format and a description of the item formats.

Table 1.1: DET Item Formats with Descriptions[1]

Item Format	Measures	Description
C-test	Reading and Writing	In the C-test, the first and last sentences are intact, while words in the sentences in the middle are "damaged" and have the second half of the word deleted. Test takers complete the damaged words in the paragraph.
Yes/No text	Reading and Writing	A set of words are presented in the yes/no text. Some of the words are actual English words, while others are pseudo-words. Test takers respond by detecting which words are real English words.
Yes/No audio	Listening and Speaking	Test takers are presented a set of words aurally. Some of them are real English words, while others are pseudo-words. Test takers respond by identifying the real English words.
Dictation	Listening and holding words (phrases) in working memory	Test takers listen to a spoken sentence or a short passage. Test takers respond by transcribing it using a computer keyboard.
Elicited speech	Reading and Speaking	Test takers are presented with a text-based sentence or short passage. Test takers respond by reading it aloud into the computer microphone.
Extended speaking picture description	Speaking	Test takers are presented a picture and are asked to describe it. Test takers respond by speaking and describing the picture.
Extended speaking aural question	Speaking	Test takers are presented aurally with a question. Test takers listen to the question, and then they respond by speaking and providing a relevant answer.
Extended speaking independent text	Speaking	Test takers are asked to explain or describe a topic. The written prompt frequently has multiple questions. Test takers respond by speaking and providing information on the topic.
Extended writing picture description	Writing	Test takers are provided a picture and are asked to describe the image. Test takers respond by writing one or more sentences describing the image.
Extended writing independent text	Writing	Test takers are provided a short statement and are asked to respond to it in at least 50 words. Test takers respond by writing a response to the statement.

[1] Item format descriptions are taken from the DET Technical Manual (LaFlair & Settles, 2020).

At the end of testing, test takers select from two topics on which to provide a more extensive writing sample. Test takers also select from three topics on which to record an oral interview. Test takers' responses to the end of testing exercises are not used to calculate DET total scores or subscores, but the responses (writing sample and interview video recording) are sent to institutions as part of the results report institutions receive.

Back to Table of Contents

Section 8: Evidence Regarding Internal Structure

Question: Does the internal structure of the DET provide support for score interpretations?

Standard 1.13: If the rationale for a test score interpretation for a given use depends on premises about the relationships among test items or among parts of the test, evidence concerning the internal structure of the test should be provided (AERA, APA, & NCME, 2014, pp. 26–27).

Standard 1.14: When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretations should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given (AERA, APA, & NCME, 2014, p. 27).

DET and Internal Structure

The DET uses 10 different item types to assess English skills. Each different item type measures various aspects of language proficiency. See the preceding section on the item formats and their descriptions to learn more about each of these item types.

To analyze the DET structure and the relationship of the item types to the reported subscores (Literacy, Conversation, Comprehension, and Production), the DET conducted non-metric multidimensional scaling analyses. Multidimensional scaling is used to investigate the validity and interpretation of subscores. Results of the analyses reveal how similar or different the questions are from each other based on test-taker response patterns on those questions. The analyses allow reducing the test questions into a smaller number (two) dimensions to examine the relationships among item types (LaFlair & Tousignant, 2020).

Results indicate that the questions work together to assess integrated modalities of language. For instance, literacy measures understanding and producing written language; conversation measures understanding and producing spoken language; comprehension measures understanding spoken and written language; production measures producing spoken and written language. As LaFlair and Tousignant demonstrate, the DET provides evidence that supports the interpretation of subscores. The authors provide a rationale supporting the interpretation of subscores and the associated interpretations of subscores.

Back to Table of Contents

Section 9: Evidence Regarding Relationships with Other Variables

Question: What is the relationship of DET scores to other related variables? Does the strength of these relationships provide support for score interpretations?

"Relationships between test scores and other measures intended to assess the same or similar constructs provide convergent evidence...Evidence of relations with other variables can involve experimental as well as correlational evidence" (AERA, APA, & NCME, 2014, pp. 16–17).

"Evidence of the relation of test scores to a relevant criterion may be expressed in various ways, but the fundamental question is always, how accurately do test scores predict criterion performance? The degree of accuracy and the score range within which accuracy is needed depends on the purpose for which the test is used" (AERA, APA, & NCME, 2014, p. 17).

Criterion Validity Evidence of DET Constructed Response Scores
Certain DET test items require test takers to construct their own responses in both writing and speaking. The item types require test takers to write a short essay responding to photo prompts and text prompts. The item types also require test takers to provide a spoken response to photo prompts, audio prompts, and text prompts. DET uses machine scoring derived from scoring rubrics and initial human scorers to rate responses.

Index for agreement between human and machine ratings: Cohen's κ (Cohen, 1960)
Cohen's Kappa estimates the extent to which different raters provide the same score to the same writing or speaking response. In the case of the DET, it is also used to determine the consistency level of human to machine scores. It is a widely accepted measure of inter-rater consistency that adjusts for chance agreement. Cohen's Kappa accounts for the possibility that raters may guess on at least some variables due to uncertainty. Estimates of Cohen's Kappa tend to be lower than traditional reliability estimates. Landis & Koch (1977) developed an interpretative framework for Cohen's Kappa. Estimates from 0.61 to 0.80 indicate substantial agreement; estimates from 0.81 to 1.00 indicate near-perfect agreement.

- The DET assessed human–machine rating agreement in a 2018 study. The reported Kappa estimates for the extended writing and speaking responses **support the validity of DET's machine scoring algorithm**. All estimates indicate substantial agreement, and they indicate that the **machine scoring algorithm is achieving appropriate levels of**

score consistency with human raters, indicating a reflection of the construct as operationalized in the scoring rubrics.

- Estimates of inter-rater consistency for human-to-machine scoring are **slightly better than the estimates for human-to-human scoring**. In the case of speaking responses, inter-rater consistency for human–machine is much greater than for human–human consistency.

Table 2.6: *Inter-rater consistency estimates for DET writing and speaking responses*

DET Extended Writing Responses		DET Extended Speaking Responses	
	Cohen's κ		Cohen's κ
Human–Human	0.77	Human–Human	0.68
Human–Machine	0.79	Human–Machine	0.77

Correlations with other measures of the same construct

The DET has also assessed the correlation of automated scores on speaking and writing items with relevant subscores from other high-stakes English proficiency tests, using self-reported test-taker scores on TOEFL and IELTS (Cardwell et al., 2021). The moderate-to-strong correlations are comparable to those reported between TOEFL and IELTS subscores (Educational Testing Service, 2010) and suggest that the DET automated writing and speaking scores measure a construct similar to that of the TOEFL and IELTS writing and speaking subscores.

Table 2.7: *Correlations of DET automated speaking and writing grades with relevant subscores of other tests*

Automated grade ↔ Criterion score	Pearson correlation	Range-corrected correlation
Writing & TOEFL writing	0.53	0.59
Writing & IELTS writing	0.42	0.47
Speaking & TOEFL speaking	0.60	0.64
Speaking & IELTS speaking	0.54	0.59

The Relationship of DET Reported Scores to Other Significant Variables

Relationship to Other Tests of English Proficiency. DET analyzed the relationship of DET scores to scores achieved on other tests of English language proficiency. They conducted a study in 2019 using test takers' self-reported scores on either the TOEFL iBT or the IELTS Academic. A total of 2,391 test takers reported scores from the TOEFL iBT, and 991 test takers reported scores from the IELTS Academic. They found a strong relationship of scores achieved on the DET to scores achieved on the TOEFL iBT and the IELTS Academic (TOEFL iBT, r = 0.74; IELTS Academic, r = 0.75; Settles et al., 2020). These correlations were nearly identical to the correlation reported by ETS between TOEFL iBT and IELTS Academic (r = 0.73). The strong correlations of DET with TOEFL iBT and IELTS Academic provide convergent evidence supporting the use of DET scores for evaluating the English language proficiency level of students applying to English-medium academic programs.

Predicting Performance. Ishikawa et al. (2016) conducted a two-year study designed to learn whether DET scores provided appropriate information regarding the academic English ability of international students at DePauw University. Faculty members rated English language ability for 85 incoming international students based on written assessments (writing ability) and interviews (oral comprehension). The findings indicated that scores on the DET predicted faculty evaluations of writing ability and oral comprehension (Writing Ability, r = 0.62; Oral Comprehension, r = 0.49). These moderate to strong correlations indicate that the DET is appropriate for predicting the English language proficiency of incoming international students.

In combination, these studies provide evidence that the DET measures English language proficiency for academic purposes, and that it is predictive of international students' readiness for academic studies in English-medium programs.

Back to Table of Contents

Section 10: Evidence Regarding the Consequences of Testing

Question: Do consequences stemming from use of the DET scores in decision-making provide support for score interpretations?

Determining whether a test is doing what it was designed and intended to do involves understanding the consequences of using scores to inform decisions. Each test-score use is associated with consequences for one or more stakeholders. Positive consequences generally are the intended outcome of test use. However, unintended consequences may result in negative impacts for stakeholders. Consequently, the *Standards* conclude that "unintended consequences merit close examination" (AERA, APA, & NCME, 2014, p. 19). The *Standards* also maintain that:

"Tests are commonly administered in the expectation that some benefit will be realized from the interpretation and use of the scores intended by the test developers" (AERA, APA, & NCME, 2014, p. 19).

"Test score interpretations for a given use may result in unintended consequences. A key distinction is between consequences that result from a source of error in the intended test score interpretation for a given use and consequences that do not result from error in test score interpretation" (AERA, APA, & NCME, 2014, p. 20).

DET Intended Consequences of Testing

Interpreting DET scores provides educational institutions and academic programs useful and accurate information differentiating L2 English students who possess the English proficiency skills necessary to succeed in an English-medium program from those L2 students who do not possess the necessary English proficiency skills. The DET positive consequence statement of testing is essentially the test program's primary claim. The validity documentation and evidence supporting this consequence is provided through

- The articulation of score-based interpretations and uses (Section 4)
- Construct definitions (Section 6)
- Evidence related to test content (Section 7)
- Evidence related to the internal structure of the test (Section 8)
- Evidence of DET score relationships to other variables (Section 9)

DET Unintended Consequences of Testing

Unintended consequences often involve scoring issues and test-taker subgroups. These unintended consequences thereby involve questions of fairness. For example, in the case of an English language proficiency test, scores could vary based on different L1 subgroups. A subgroup whose L1 is more closely related to English may score higher because English is easier for the subgroup members to learn as compared to test takers who speak a language that is further removed or distant from English. If this is the case, the higher scores accurately reflect higher English proficiency, and so this discrepancy is not indicative of unfairness. However, other aspects of testing and scoring might result in construct-irrelevant variance, which may disadvantage those whose language is further removed from English. If the test has writing tasks but the scoring system cannot recognize non-English characters and punctuation, the resulting scoring penalties may result in construct-irrelevant variance, and as a result may disadvantage speakers of languages written with such characters if they inadvertently include them in their writing. In turn, this scoring approach may raise questions of fairness. The scoring process produces unintended consequences with one subgroup performing better than the other due to their typing ability or different writing conventions. The emergence of such potential unintended consequences generates the need to critically evaluate the testing and scoring processes, carefully evaluating them to ensure that all groups are being treated fairly and no unintended consequences are being introduced.

DET is embarking on a research agenda to evaluate all components of fairness as related to the testing and scoring process.

Back to Table of Contents

Appendix A: Can-Do Statements for the DET

DET scores 10 to 55

- Can understand very basic English words and phrases.
- Can understand straightforward information and express themselves in familiar contexts.

DET scores 60 to 85

- Can understand the main points of concrete speech or writing on routine matters such as work and school.
- Can describe experiences, ambitions, opinions, and plans, although with some awkwardness or hesitation.

DET scores 90 to 115

- Can fulfill most communication goals, even on unfamiliar topics.
- Can understand the main ideas of both concrete and abstract writing.
- Can interact with proficient speakers fairly easily.

DET scores 120 to 160

- Can understand a variety of demanding written and spoken language including some specialized language use situations.
- Can grasp implicit, figurative, pragmatic, and idiomatic language.
- Can use language flexibly and effectively for most social, academic, and professional purposes.

Back to Table of Contents

Appendix B: DET Content Framework References

The appendix contains a list of sources applied by DET designers to develop the framework for language proficiency: reading, writing, speaking, and listening.

Abraham, R. G., & Chapelle, C. A. (1992). The meaning of cloze test scores: An item difficulty perspective. *The Modern Language Journal, 76*(4), 468–479.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge, UK: Cambridge University Press.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning, 42*, 529–555.

Bachman, L. F., & Palmer, A. S. (19960. *Language testing in practice*. Oxford, UK: Oxford University Press.

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

Beinborn, L., Zesch, T., & Gurevych, I. (2014). Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics, 2*, 517–530.

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

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

Capel, A. (2010). A1-B2 vocabulary: Insights and issues arising from the English Profile Wordlists project. *English Profile Journal, 1*(1) 1–11.

Capel, A. (2012). Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal, 3*(1) 1–14.

Coombe, C. (2018). An A to Z of Second Language Assessment: How Language Teachers Understand Assessment Concepts. London, UK: British Council.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge, UK: Cambridge University Press.

Council of Europe (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors*. Strasbourg: Council of Europe. URL: https://rm.coe.int/cefr-companion-volumewith-new-descriptors-2018/1680787989. (accessed on 16 July 2021).

Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Strasbourg: Council of Europe. URL: https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4. (accessed on 16 July 2021).

Culligan, B. (2015). A comparison of three text formats to assess word difficulty. *Language Testing, 32*(4), 503–520.

Cumming, A. (2013). Assessing integrated skills. In A. J. Kunnan (Ed.), *The companion to language assessment, 1,* 216–229. Hoboken, NJ: Wiley-Blackwell.

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

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from Four L1s. *Studies in Second Language Acquisition, 19*(1) 1–16.

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning, 48,* 393–410.

DuBay, W. H. (2006). *Smart language: Readers, readability, and the grading of text.* Costa Mesa, CA: Impact Information.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly, 39*, 399–423.

Freedle, R., & Kostin, I. (1993). *The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items*. ETS Research Report 93-13. Princeton, NJ: ETS.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*, 201–223.

Hinkel, E. (2006). Current perspectives on teaching the four skills. *TESOL Quarterly, 40*(1), 109–131.

Hinkel, E. (2010) Integrating the four skills: Current and historical perspectives. In R. B. Kaplan (Ed.), *Oxford Handbook in Applied Linguistics* 2nd ed., (pp. 110–126). Oxford, UK: Oxford University Press.

Isbell, D. (2017). Assessing C2 writing ability on the Certificate of English Language Proficiency: Rater and examinee age effects. *Assessing Writing, 34*, 37–49.

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

Lin, W. Y., Yuan, H. C., & Feng, H. P. (2008). Language reduced redundancy tests: A re-examination of cloze and c-test. *Journal of Pan-Pacific Association of Applied Linguistics, 12*, 61–79.

Litman, D., Strik, H., & Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly, 15*, 294–308.

Loukina, A., Yoon, S. Y., Sakano, J., Wei, Y., & Sheehan, K. (2016, December). Textual complexity as a predictor of listening items in language proficiency tests. In *Proceedings of the International Conference on Computational Linguistics (COLING)* (pp. 3245–3253) Osaka, Japan. Retrieved from https://www.aclweb.org/anthology/C16-1306.pdf

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research* (pp. 211–232). Modena, Italy: Eurosla Monographs.

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. In R. Chacon-Beltran, C. Abello-Contesse, & M. Torreblanca-Lopez (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83–98). Bristol, UK: Multilingual Matters.

Munro, M.J., & Derwing, T.M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*, 73–97.

Perez-Beltrachini, L., Gardent, C., & Kruszewski, G. (2012, June). Generating grammar exercises. In *Proceedings of the 7th Workshop on Building Educational Applications Using Natural Language Processing* (pp. 147–156) Montreal, Canada. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.935.1408&rep=rep1&type=pdf

Reichert, M., Keller, U., & Martin, R. (2010). The c-test, the TCF and the CEFR: A validation study. In R. Grotjan (Ed.), *Der C-test: Beitragen aus der aktullen Forschung/The C-test: Contributions from current research* (pp. 205–231). Frankfurt am Main, Germany: Peter Lang.

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

Sung, Y. T., Lin, W. C., Dyson, S. B., Change, K. E., & Chen, Y. C. (2015). Leveling L2 texts through readability: Combining multilevel linguistic features with CEFR. *The Modern Language Journal, 99*(2), 371–391.

Susanti, Y., Nishikawa, H., Tokunaga, T., & Obari, H. (2016). Item difficulty analysis of English vocabulary questions. In *Proceedings of the 8th International Conference on Computer Supported Education (CSEDU)* (pp. 267–274). Retrieved from https://www.cl.c.titech.ac.jp/_media/publication/susanti_2016aa.pdf

Vajjala, S., Meurers, D., Eitel, A., & Scheiter, K. (2016, December). Towards ground computational linguistic approaches to readability: Modeling reader–text interaction for easy and difficult texts. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)* (pp. 38–48) Osaka, Japan. Retrieved from https://www.aclweb.org/anthology/W16-4105/

Vajjala, S., & Rama, T. (2018, June). Experiments with universal CEFR classification. In *Proceedings of the 13th Workshop on Innovative Use of Natural Language Processing for Building Educational Applications* (pp. 147–153) New Orleans, LA. Retrieved from https://www.aclweb.org/anthology/W18-0515.pdf

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

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

Widdowson, H. G. (1978). *Teaching language as communication.* Oxford, UK: Oxford University Press.

Xia, M., Kochmar, E., & Briscoe, T. (2016, June). Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Building Educational Applications Using Natural Language Processing* (pp. 12–22) San Diego, CA. Retrieved from https://www.aclweb.org/anthology/W16-0502/

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.

Back to Table of Contents

References

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

Brenzel, J., & Settles, B. (2017). *The Duolingo English Test – Design, validity, and value*. Retrieved from https://s3.amazonaws.com/duolingo-papers/other/DET_ShortPaper.pdf.

Byram, M., & Parmenter, L. (Eds.). (2012). *The common European framework of reference: The globalisation of language education policy*. Bristol, UK: Multilingual Matters.

Cardwell, R. L., LaFlair, G. T., & Settles, B. (2021). *Duolingo English Test: Technical Manual*. Retrieved from: https://duolingo-papers.s3.amazonaws.com/other/det-technical-manual-current.pdf

Council of Europe. (2001). *Common European Framework of References for Languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.

Educational Testing Service. (2010). *Linking TOEFL iBT scores to IELTS scores—A research report*. Educational Testing Service Princeton, NJ.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practices, 23*(1), 17–27.

Ishikawa, L., Hall, K., & Settles, B. (2016). *The Duolingo English Test and Academic English*. Retrieved from: https://s3.amazonaws.com/duolingo-papers/reports/DRR-16-01.pdf.

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

LaFlair, G. T. (2020). *Duolingo English Test: Subscores*. Retrieved from: https://duolingo-papers.s3.amazonaws.com/reports/subscore-whitepaper.pdf

LaFlair, G. T., & Settles, B. (2020). *Duolingo English Test: Technical Manual*. Retrieved from: https://duolingo-papers.s3.amazonaws.com/other/det-technical-manual-current.pdf

LaFlair, G. T., & Tousignant, J. (2020). *Subscores: Improving how we report Duolingo English test results*.

https://blog.duolingo.com/subscores-improving-how-we-report-duolingo-english-test-results-2/

Moritoshi, P. (2001). *The Test of English for International Communication (TOEIC): Necessity, proficiency levels, test score utilization, and accuracy*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.199.475&rep=rep1&type=pdf

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

Von Davier, A. (2020). *Test score comparability for digital-first assessments in a computational psychometrics framework*. Duolingo, Inc.

Back to Table of Contents