Analysis of the Scoring and Reliability for the Duolingo English Test

Table of Contents

Section 1: Purpose of this Document

The purpose of this document is to report our review of the research supporting the Duolingo English Test (DET) scoring and score reliability/precision. We used the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) and DET's documentation to analyze the topics identified below. As the *Standards* address scoring and reliability issues across various chapters, this document integrates sections from chapters 2, 5, and 6. Key considerations related to scoring and reliability associated with the chapters include:

> Standard 2.0: Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use (AERA, APA, & NCME, 2014, p. 42).

> Standard 5.0: Test scores should be derived in a way that supports the interpretations of test scores for the proposed uses of tests. Test developers and users should document evidence of fairness, reliability, and validity of test scores for their proposed use (AERA, APA, & NCME, 2014, p. 102).

> Standard 6.0: To support useful interpretations of score results, assessment instruments should have established procedures for test administration, scoring, reporting, and interpretation. Those responsible for administering, scoring, reporting, and interpreting should have sufficient training and supports to help them follow the established procedures. Adherence to the established procedures should be monitored, and any material errors should be documented and, if possible, corrected (AERA, APA, & NCME, 2014, p. 114).

Questions Analyzed

We analyzed the following questions related to the scoring of the DET:

1. What are the established scoring procedures for the DET?
    - What are DET's procedures for generating items and estimating item-level difficulty?
    - What are DET's procedures for selecting items for administration?
    - What are DET's procedures for scoring item responses?
    - What are DET's procedures for deriving total scores and subscores?

- What are DET's procedures for supporting valid and appropriate score interpretations?

2. What is the level of reliability/precision of the DET scores?
    - What is the reliability/precision of the DET total scores?
    - What is the reliability/precision of the DET subscores?
    - What is the amount of error contained within the DET scores?

In what follows, we summarize our analysis and use examples of evidence collected from DET's documentation to address the questions.

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: DET Scoring Procedures

> Standard 5.16: When test scores are based on model-based psychometric procedures, such as those used in computerized adaptive or multistage testing, documentation should be provided to indicate that scores have comparable meaning over alternate sets of test items (AERA, APA, & NCME, 2014, p.106).

Questions:

- **What are the established scoring procedures for the DET?**
- **What are the DET's procedures for selecting items for administration?**

In this section, we begin with general findings for DET's item generation procedures and resulting test scores. Then, we summarize research conducted on the DET related to the above-mentioned questions.

Overview of Scoring Procedures

The DET uses machine learning and natural language processing models to create proficiency scales. Then, these linguistic models are used to estimate item difficulty for use with computer-adaptive test (CAT) algorithms. As a result, the need for pilot testing with human subjects is obviated. The use of these models and algorithms produce:

- a large pool of items that satisfies test security requirements
- scores that have high levels of reliability/precision
- scores that are highly correlated with scores achieved on other high-stakes English assessments (i.e., TOEFL iBT and IELTS)

Computer-Adaptive Testing (CAT)

Computer-adaptive testing (CAT) (Segall, 2005; Wainer, 2000) is used to determine the items that test takers respond to. As compared to fixed-form tests such as paper-and-pencil tests, CAT enables shorter tests, uniformly precise scores, and greater item security (Brenzel & Settles, 2017; Maris, 2020; Settles, LaFlair, & Hagiwara, 2020).

Item responses are automatically scored on a continuous zero–one scale (Settles et al., 2020). DET's CAT administration is built on five item formats (C-test, audio yes/no vocabulary, text yes/no vocabulary, dictation, and elicited imitation) and draws on a bank of more than 25,000 test items. DET indexed items into eleven bins for each format. Each bin corresponds to a range on the item difficulty scale; Settles et al., 2020).

The CAT administration algorithm randomly chooses the first item format to use, and then cycles through the remaining item formats to determine the format for each subsequent item. All five formats have equal representation.

Each item format has its own machine learning– and natural language processing–based statistical grading procedure. (See Section 6: Scoring Item Responses below for greater detail on the specifics of grading each item format). Following the administration of the first four items, the CAT algorithm estimates a provisional score. The estimated provisional score then determines which bin the next item is drawn from. After the test taker responds to the new item, the CAT re-estimates the score and selects the next item from the bin aligned to the test taker's estimated provisional score. The result of the CAT process is that, for test takers who are struggling, they respond to more basic items; for test takers who are doing well, they respond to more challenging items. The goal of the CAT process is to estimate test takers' abilities as precisely as possible with as few items as possible (Settles et al., 2020).

The CAT process continues until the test exceeds 25 items or 40 minutes in length (Settles et al., 2020). Test takers respond to a minimum of three items of each format and a maximum of seven items of each format, with a median rate of six items of each format (LaFlair & Settles, 2020).

In addition to the five CAT item formats, test takers receive eight extended-response speaking and writing tasks, which present a picture or verbal prompt that test takers must respond to either orally or in writing, depending on the task. These items are categorized into three difficulty levels and are chosen based on the test taker's performance on the items in the CAT portion. Responses to extended-response items are scored using an automatic scoring model unique to each task type.

Once items and tests are administered using the pre-calibrated item bank, actual response data under high-stakes testing conditions are available. The data can be used to (a) update the predetermined item difficulties where needed, (b) evaluate the fit of the item response theory model, and (c) further train the machine learning and natural language processing algorithms to obtain more accurate pre-calibrated item difficulties.

Back to Table of Contents

Section 4: Score Reliability/Precision – Overview

Score Reliability refers to the **consistency** and **precision** of test scores.

Standard 2.0: appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use (AERA, APA, & NCME, 2014, p. 42).

- "The need for precision **increases** as the **consequences of decisions and interpretations** grow in importance" (AERA, APA, & NCME, 2014, p. 33).

- "Reliability/precision of data ultimately bears on the **generalizability** or **dependability** of the scores and/or the **consistency of classification** of individuals derived from the score" (AERA, APA, & NCME, 2014, p. 34).

- When a test includes constructed response items, evaluating reliability/precision requires analyzing the **consistency of the scoring process**.

Analyzing and Evaluating Reliability

Chapter 2 of the current *Standards* (AERA, APA, & NCME, 2014) provides guidance and standards for conceptualizing and studying reliability.

- In reliability studies, scores are analyzed and interpreted as to whether they provide appropriate levels of precision.
- Reliability estimates convey the proportion of test score variance that is attributable to true differences in the intended test construct, and the proportion of variance that is attributable to measurement error.
- Whether an appropriate level of precision has been achieved is evaluated in the context of the decisions/consequences emanating from the test score interpretation.
- The type of reliability study conducted should be determined by the structure of the test and the associated decisions or consequences.

Psychometricians have developed various approaches for studying and evaluating score reliability and precision. These approaches include classical test theory and item response theory.

1. Classical Test Theory (CTT)
    a. The use of classical test theory has led to the development of reliability coefficients to represent true score variance to total score variance.
    b. Coefficients of internal consistency indicate the reliability of results within an assessment and address the question: Are test takers responding consistently across a set of items/tasks?
    c. Coefficients of external reliability indicate the consistency of results across testing occasions and address the question: Are test takers' scores consistent across different testing occasions?
    d. Generalizability Theory analyzes score variance and the sources of variance and addresses the question: What proportion of score variance is attributable to differences in test takers on the intended construct?
2. Item Response Theory (IRT)
    a. IRT analyzes test information across the range of abilities (score distribution).
    b. IRT allows test developers to maximize score information at key points in the score distribution (cut scores), making decisions derived from the scores more precise.
    c. The use of IRT enables psychometricians to examine how much information a set of assessment tasks provides at a specific ability level or score.

In selecting an approach for studying reliability, researchers must first analyze how test scores are interpreted and the consequences stemming from the interpretations. Following the articulation of score interpretations and consequences, they should evaluate the structure of the test (e.g., item types, test-taker responses, administration). Based on an integration of these conceptualizations, the researcher should select an approach such as classical test theory or IRT to analyze reliability that is best aligned to the interpretations and consequences.

Score Reliability and its Connection to Validity

Although score reliability is generally analyzed and considered independently from validity, score reliability and score precision has implications for the validity of score-based interpretations. As the *Standards* suggest:

"To the extent that scores are not consistent across replications of the testing procedure, their potential for accurate prediction of criteria, for beneficial examinee diagnosis, and for wise decision making is limited" (AERA, APA, & NCME, 2014, pp. 34-35).

The consistency of scores can be estimated through the use of the Standard Error of Measurement (SEM).

Standard Error of Measurement

The estimated SEM "is an indicator of a lack of consistency in the scores generated by the testing procedure for some population. A **relatively large SEM indicates relatively low reliability/precision**" (AERA, APA, & NCME, 2014, p. 34).

Therefore, one goal is to have a low SEM across the score range. A low SEM indicates a more precise measurement of student proficiency, warranting increased confidence in the test scores.

Evaluating Reliability Studies

Establishing criteria for appropriate levels of score reliability is difficult without also considering the testing context and score use (e.g., the construct, population of test takers, types of decisions stemming from test scores). Despite these obstacles, psychometricians and researchers have found that reliability coefficients for test scores measuring cognitive constructs with consequential interpretations should reach 0.80 or above. Reliability coefficients around 0.90 in this context are interpreted as excellent (U.S. Department of Labor, 2006).

Reliability Coefficients for the DET

Test–retest reliability is reported at 0.90. This value is **high** for a computer-adaptive test. Reliability coefficients typically range from 0.8 to 0.9 for established standardized tests using identical forms. Scores from CAT's are often found to be less reliable because the items vary from test session to test session. Additional information regarding DET's reliability coefficients are provided in later sections in this report.

Back to Table of Contents

Section 5: Procedures for Item Difficulty Estimation

Question: What are DET's procedures for generating items and estimating item-level difficulty?

> Standard 5.16: When test scores are based on model-based psychometric procedures, such as those used in computerized adaptive or multistage testing, documentation should be provided to indicate that scores have comparable meaning over alternate sets of test items (AERA, APA, & NCME, 2014, p. 106).

The DET contains 10 scored item formats. Five item formats (C-test, audio yes/no vocabulary, text yes/no vocabulary, dictation, and elicited imitation) are delivered using a CAT algorithm. The other five item formats are open-ended speaking and writing tasks. Test takers receive a minimum of three and a maximum of seven of each of the CAT items; they receive four each of the open-ended writing (two formats) and speaking tasks (three formats). The difficulty of all CAT items is predicted using one of two ML–/NLP–based models—the passage model or the vocabulary model.

The Passage Model

DET needed to develop an accurate method for automatically aligning authentic passages to the six CEFR levels. Accurate classification of passages was needed for C-test, dictation, and elicited imitation content. To this end, the DET developers created the passage model by applying the principles of natural language processing and machine learning. DET created the passage model by leveraging a variety of corpora from online resources, using a combination of ranking and regression techniques to train the passage models to predict difficulty of multi-word texts.

Given the paucity of available model training sets to profile CEFR texts or discourse features, a semi-supervised learning approach was applied (Zhu & Goldberg, 2009). First, passages were ranked by overall difficulty. Then, the CEFR levels were propagated from a small number of labeled texts to many more unlabeled texts that are similar in difficulty. Moreover, word length and sentence length features were used to train a word-level unigram language model to produce log-likelihood and Fisher score features (similar to a weighted bag of words).

An initial training corpus was used. The corpus was based on online English language self-study websites (e.g., free test preparation resources for popular English proficiency exams) consisting of reference phrases and texts from reading comprehension exercises, all organized by CEFR level. The documents were segmented and CEFR labels were propagated down to the paragraph level, which resulted in 3,049 CEFR-labeled passages. Due to the small size of the CEFR corpus, passage pairs from English Wikipedia and Simple Wikipedia English were used to supplement the CEFR corpus. These materials were further supplemented with thousands of sentences downloaded from a crowd-sourced self-study English language database (tatoeba.org). Using these resources, ranking experiments were conducted using standard linear regression methods. DET found that weighted-softmax regression provided the best model fit. They cross-validated the findings analyzing 3,049 passages from the CEFR corpus and found that the model provided strong predictive power. The cross validation indicated a strong correlation of the expert CEFR judgment and the model predicted level (r = 0.76). DET commented that it produced a conservative estimate, "since we propagate CEFR document labels down to paragraphs for training and evaluation and this likely introduces noise (e.g., C1-level articles may well contain A2-level paragraphs)" (Settles et al., 2020, p. 254).

DET continued its evaluation with a post-hoc validation. In the post-hoc study, the DET employed four experts in linguistics to compose 2,349 passages, written to the six CEFR levels. The passages were then converted to C-test passages. Each passage was written by an expert and vetted independently by a second expert, with both experts needing to agree on the final CEFR level. The passages were then rated by the model. The relationship of expert ratings to model ratings was strong to moderate. Moreover, there was a flattening of the linear relationship between levels C1 and C2. The flattening indicated that both the experts and the model struggled to distinguish between fine-grain differences of the two highest levels of linguistic complexity.

The results of both the initial validation study and the post-hoc study provide support for DET's model for classifying passages for use in the C-test, dictation, and elicited imitation tasks (Maris, 2020). Additionally, DET's findings illustrate the application of machine learning to accurately classify written text materials for use in testing.

The Vocabulary Model

The vocabulary model is analogous to the passage model but for individual words, and was developed based on a standard setting exercise conducted with a panel of experts who had doctoral degrees in linguistics and ESL teaching experience. Based on previous work (Capel, 2010, 2012; Cambridge English, 2012), subject matter experts created an English CEFR vocabulary wordlist (i.e., a dictionary of 6,823 English words labeled by CEFR level primarily in the B1-B2 range). The labeled wordlist was used to train a vocabulary model to assign difficulty levels to words based on features that could also be computed for pseudo-words. More advanced words are rarer and generally have Graeco-Latin etymologies, whereas more basic words are common and generally have Anglo-Saxon origins. While the pseudowords do not actually exist, their properties appear to hold (e.g., “cload” seems more Anglo Saxon and more common than “fortheric” should be). These illustrations suggest that the model is capturing qualitative subtleties in the English lexicon, as they relate to proficiency levels (Settles et al., 2020). The vocabulary model is used to determine the difficulty of audio and text yes/no vocabulary items.

Each of the item formats is described next along with a description of how the items are scored. Items are graded automatically using statistical procedures developed specifically for each item format.

Back to Table of Contents

Section 6: Procedures for Scoring Item Responses

Question: What are DET's procedures for scoring item responses?
This section describes DET's procedures for scoring item responses for each item format.

Scoring the C-test, Dictation, and Elicited Imitation Tasks
C-test. The C-test items consist of short language excerpts taken from authentic texts. In the C-test items, the first and last sentences remain fully intact and unchanged, while words in the remaining text are 'damaged.' The sentences are damaged by removing the second half of every other word. Test takers respond by completing the damaged words. The premise of the C-test is that language processing requires activation of more than one language component (e.g., grammar and vocabulary) and more than one skill (e.g., reading and comprehension).

C-test responses are scored by aligning responses against expected reference texts. Similarities and differences are evaluated. The response is graded using a weighted average of the correctly filled word-gaps. The weight of each gap is proportional to its length in characters, with longer gaps being weighted more heavily. A probabilistic grade based on a binary logistic regression model is applied to determine the item score.

Dictation. The dictation tasks require test takers to listen to a spoken sentence or short passage, and then transcribe it using the computer keyboard. Test takers have one minute to listen to and transcribe the spoken statement. They may listen to the statement up to three times. The difficulty of the dictation task is determined by its lexical complexity and length. This task assesses test takers' ability to recognize individual words and to hold them in memory long enough to reproduce them. The task measures listening comprehension and writing skill.

Dictation responses are graded using logistic regression classifiers. Test takers' written submissions are aligned to an expected reference text. Features representing the differences in alignment are extracted. Models trained on aggregate human judgments of correctness for tens of thousands of test item submissions are used to produce probabilistic training labels, which describe the probability that a random English-speaker would find a particular transcription faithful and accurate.

Elicited imitation. The elicited imitation tasks measure reading and speaking. The tasks require test takers to read a complete written sentence aloud. Test takers respond by using the computer's microphone to record themselves speaking the written sentence. The goal is to evaluate the accuracy and intelligibility of the speech production. To evaluate the general clarity of the speech, DET technology extracts features of spoken language, including acoustic and fluency features.

Responses to the elicited imitation tasks are graded using logistic regression classifiers. The responses are automatically transcribed using speech recognition. Then, they are aligned to an expected text. Features representing the differences in alignment are extracted. Models trained on aggregate human judgments of accuracy and intelligibility for tens of thousands of task submissions are used to produce probabilistic training labels, which describe the probability that a random English-speaker would find a particular utterance faithful, intelligible, and accurate.

Scoring the Audio and Text Yes/No Vocabulary Tasks

In the Audio and Text Yes/No Vocabulary Tasks, test takers are presented with a collection of English words and pseudo-words (in two separate modalities – audio and text). Test takers are asked to discriminate between the English words and the pseudo-words. These tasks measure spoken and written vocabulary, which are required to meet communication needs across the CEFR levels.

The **yes/no vocabulary** format is graded using a sensitivity index. A sensitivity index is a measure of separation between signal (word) and noise (pseudo-word) distributions (Beeckmans, Eyckmans, Janssens, Dufranne, & Van de Velde, 2001). This grading can be interpreted as the probability that test takers can discriminate between English words and pseudo-words at various levels of difficulty.

Scoring of Writing and Speaking Tasks

The speaking tasks include three different prompt types (picture description, text-based, and audio). The writing tasks include two prompt types (picture description and text-based).

DET's experts in machine learning and natural language processing developed the **automated scoring algorithms** for the writing and speaking tasks. Separate algorithms are used; one for the writing tasks and the other for the speaking tasks.

The speaking and writing scoring systems evaluate each task based on the following features:

- Grammatical accuracy

- Grammatical complexity
- Lexical sophistication
- Lexical diversity
- Task relevance
- Length
- Fluency and acoustic features (speaking)

Regarding the evaluation of fluency and acoustic features in the extended speaking responses, DET scoring technology relies not only on computer speech recognition software, but it also is able to evaluate acoustic properties of speech like intonation, rhythm, and stress. These 'suprasegmental' properties of speech significantly contribute to its intelligibility (Brenzel & Settles, 2017).

Back to Table of Contents

Section 7: Procedures for Deriving Total Scores and Subscores

Question: How are DET total scores and subscores derived?
Once the algorithm converges, the final reported score is not the provisional MLE point-estimate used during CAT administration. Rather, for each CAT item type, the probability is computed for each possible $\theta \in [0, 10]$ and normalized into a posterior distribution in order to create a weighted average score. These weighted average scores of each CAT item type are then used with the scores of the speaking and writing tasks to compute a total score and the four subscores. All five scores are estimated independently, and they are weighted combinations of item type scores (LaFlair, 2020; LaFlair & Tousignant, 2020).

The DET reports subscores to provide information on test takers' proficiency in different components of language such as speaking, writing, reading, and listening. The subscores are provided to test takers and institutions so that they have a more complete understanding of the test takers' abilities to function in English at their institution. The subscores provide more nuanced information about a test taker's language abilities, without requiring them to take another test.

Subscores were selected to reflect an understanding that natural language use is complex, needs to integrate different components of language, and varies by situation and context. The subscores also reflect an understanding that people use multiple skills simultaneously to communicate. For instance, attending a lecture may require students' comprehension skills (i.e., listening and reading) whereas participating in study groups may require listening and speaking. Thus, to reflect natural language use, the reported subscores represent *integrated modalities*.

The DET reports subscores for **Literacy, Conversation, Comprehension,** and **Production.**
To analyze the relationship among item types, non-metric multidimensional scaling (MDS) was used. The MDS analyses were conducted to examine how similar (or different) the item types are from each other based on test takers' aggregated scores by item type. The analyses allowed reducing the test questions into a smaller set of (two) dimensions to examine the relationships among the different questions (LaFlair & Tousignant, 2020). The results shown in Table 2 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. More specifically, Table 2.1 presents the four subscores along with the item formats that are used to estimate the subscore and provides information around its measurement definition.

Table 2.1: *DET Subscores, Measures, and Item Types Contributing to Subscore Estimates*

DET Subscore	Measures	Item Type
Literacy	Understanding and producing written language	c-test, writing, and yes/no text
Comprehension	Understanding spoken and written language	c-test, dictation, elicited speech, yes/no text, yes/no audio
Conversation	Understanding and producing spoken language	speaking, dictation, elicited speech, yes/no audio
Production	Spoken and written language	speaking, writing

Back to Table of Contents

Section 8: Procedures for Supporting Valid and Appropriate Score Interpretations

Question: How does the DET score scale support valid and appropriate interpretations? "Scale scores, proficiency levels, and cut scores can be central to the use and interpretation of test scores. For that reason, their defensibility is an important consideration in test score validation for the intended purposes" (AERA, APA, & NCME, 2014, p. 95).

Standard 5.1: Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretations of scale scores, as well as their limitations (AERA, APA, & NCME, 2014, p. 102).

Standard 5.2: The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly (AERA, APA, & NCME, 2014, p. 102).

Standard 5.16: When test scores are based on model-based psychometric procedures, such as those used in computerized adaptive or multistage testing, documentation should be provided to indicate that scores have comparable meaning over alternate sets of items (AERA, APA, & NCME, 2014, p. 106).

DET Scores and CEFR Levels and Descriptions

The total English proficiency score and the four subscores are reported on a 10–160 scale that is aligned to skills in the CEFR. Scores are reported in 5-point increments. Test takers receive a score report that provides them with their score and the associated English proficiency skills. Table 2.2 provides DET scores aligned to the CEFR proficiency levels.

Table 2.2: *DET scores aligned to the CEFR proficiency levels and their skill descriptions*

DET Score	CEFR	Skill Description
10-55	A1/A2	● Can understand very basic English words and phrases.

		● Can understand straightforward information and express themselves in familiar contexts.
60-85	B1	● 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.
90-115	B2	● 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.
120-160	C1/C2	● 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.

The DET's Score Scale

A suitable score scale facilitates the understanding of score meaning, enables the identification of performance differences, and decreases the likelihood of score misinterpretations (Kolen & Brennan, 2014). The 10–160 scale combined with the alignment to the CEFR skills facilitates understanding and interpretation of the scores. The scale assists in differentiating individuals who have a beginning understanding of English from individuals who have mastered English. The scale provides sufficient score points so that differential performance levels are conveyed. Lastly, DET documentation enables users to have a sound understanding of the meaning of scores and avoid score misinterpretations. Table 2.3 presents descriptive total score and subscore statistics from test takers who took the DET between August 2020 and July 2021.

Table 2.3: *Summary statistics for the DET Test: Total Score and Subscores* (LaFlair & Settles, 2020)

Score	Mean	SD	25th Percentile	Median	75th Percentile
Literacy	107.45	20.06	95	110	120
Conversation	98.54	22.01	85	100	115
Comprehension	116.03	19.99	105	120	130
Production	85.18	22.61	70	85	100
Total	107.40	19.29	95	110	120

Back to Table of Contents

Section 9: DET Total Score Reliability

Question: What is the reliability and precision of the DET total scores?

Standard 2.3: For each **total test score**, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported (AERA, APA, & NCME, 2014, p. 43).

DET Reliability/Precision Studies

The studies conducted by the DET to analyze the reliability and precision of the DET are summarized in Table 2.4. The information provided in the table indicates:

- DET estimates of test–retest reliability coefficients range from .79 (Ye, 2014) to 0.90 (Cardwell, LaFlair, & Settles, 2021)
- DET estimates of internal consistency using split-half analyses are 0.95 (LaFlair & Settles, 2020) and 0.96 (Settles, 2020; Settles et al., 2020).
- Because the items used in computer-adaptive tests vary from one session to the next, their reliability is often lower than standardized tests with identical forms, which typically have coefficients ranging from 0.8 to 0.9 (Settles, 2016).
- These are **high reliability coefficients** for a CAT, indicating that the reliability/precision of the DET total scores is well above the acceptable range.

Table 2.4: *DET total score reliability estimates*

Study Authors & Year	Reliability*	*n*	Coefficient
Cardwell, LaFlair, and Settles (2021)	Test–retest	**	0.90
LaFlair and Settles (2020)	Test–retest	10,187	0.82
	Internal Consistency	8,041	0.95
Settles, LaFlair, and Hagiwara (2020)	Test–retest	526	0.80
	Internal Consistency	9,309	0.96
Settles (2016)	Test–retest	8,130	0.84

	Internal Consistency	8,130	0.96
	Hashing Alpha	8,130	0.93
Ye (2014)	Test–retest	107	0.79

*Split-half studies were conducted to obtain internal consistency estimates.

**The method used to adjust for the non-representative repeat tester population makes the sample size misleading and so it was not reported

Back to Table of Contents

Section 10: DET Subscore Reliability/Precision

Question: What is the reliability/precision of the DET subscores?

> Standard 2.3: For each total test score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported (AERA, APA, & NCME, 2014, p. 43).

Table 2.5 presents the reliability coefficients (test–retest and split-half reliability estimates) and Standard Error of Measurement for the four DET subscores. Because subscores are calculated based on a partial item set, the expectation is that the reliability coefficients for the subscores will be less than for the total score. As a result, the criteria for evaluating the sufficiency of the reliability coefficients for subscores is lower than for total scores.

DET Subscore Reliability/Precision Studies – Summary of Findings

- The reliability coefficients for subscores range from 0.75 to 0.95. The estimates appear to follow similar patterns in both studies. The reliability estimates for subscores appear to all fall in the range of **good to excellent** rating for subscore precision.

Table 2.5: ***Reliability coefficients for DET subscores of Literacy, Conversation, Comprehension, and Production***

Study	Subscore	Reliability	*n*	Coefficient
DET Technical Manual (LaFlair & Settles, 2020)	Literacy	Test–retest	—	0.88
		SEM	—	6.95
	Conversation	Test–retest	—	0.86
		SEM	—	8.23
	Comprehension	Test–retest	—	0.86
		SEM	—	7.48
	Production	Test–retest	—	0.86
		SEM	—	8.46
LaFlair and Tousignant (2020)	Literacy	Test–retest	47,654	0.82
		Split-half	47,654	0.89
	Conversation	Test–retest	47,654	0.80
		Split-half	47,654	0.93
	Comprehension	Test–retest	47,654	0.78
		Split-half	47,654	0.95
	Production	Test–retest	47,654	0.93
		Split-half	47,654	0.76

Back to Table of Contents

Section 11: Standard Error of Measurement

Question: What is the amount of error contained within the DET scores?

> **Standard 2.13:** The standard error of measurement, both overall and conditional (if reported), should be provided in units of each reported score (AERA, APA, & NCME, 2014, p. 45)

The Standard Error of Measurement (SEM)

- Estimates a measure of the score distribution around a test taker's true score
- Estimates a theoretical confidence interval or score range around the estimated true score
- Is more informative than reliability coefficients (because it is reported in score units)

DET SEM Total Score Estimates:

- Settles (2016) reports an **SEM for the DET Total Score of 5.5** (on the old 100-point scale)
- The DET Technical Manual reports an **SEM for the DET Total Score of 6.10** (on the revised 160-point scale)

DET Standard Error of Measurement: Summary

- The DET SEM for the revised 10–160 scale was estimated at **6.10**. The estimate indicates relatively **small error of measurement** and **satisfactory measurement reliability/precision**.
- The DET CSEM for the 100-point scale provided **estimated errors of around six points in the areas of the distribution** where most test takers scored. A six-point score interval for the majority of test takers provides an **acceptable level** of score reliability/precision.

Back to Table of Contents

References

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

Beeckmans, L. F., 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. Retrieved from https://journals.sagepub.com/doi/10.1177/026553220101800301.

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.

Cambridge English. (2012). Preliminary wordlist. Retrieved from https://www.cambridgeenglish.org/images/84669-pet-vocabulary-list.pdf.

Capel, A. (2010). A1-B2 vocabulary: Insights and issues arising from the English Profile Wordlists project. *English Profile Journal, 1*. Retrieved from https://www.cambridge.org/core/journals/english-profile-journal/article/a1b2-vocabulary-insights-and-issues-arising-from-the-english-profile-wordlists-project/E57847F6C5574124B2354F9BEEC005FA/core-reader.

Capel, A. (2012). Completing the English vocabulary profile: C1 and C2 vocabulary. *English Profile Journal, 3*. Retrieved from https://www.cambridge.org/core/journals/english-profile-journal/article/completing-the-english-vocabulary-profile-c1-and-c2-vocabulary/418955FC7ED2455E98A499BC40C2C816

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*, 37–46.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practice* (3rd Ed.). New York, NY: Springer.

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/

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Maris, G. (2020) *The Duolingo English test: Psychometric considerations.* https://duolingo-papers.s3.amazonaws.com/reports/DRR-20-02.pdf

Segal, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement* (pp. 429–438). New York, NY: Academic Press.

Settles, B. (2016). *The reliability of the Duolingo English test.* Retrieved from: https://s3.amazonaws.com/duolingo-papers/reports/DRR-16-02.pdf

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

The Duolingo English Test: Security, Proctoring, and Accommodations Security, Proctoring, and Accommodations (2020). englishtest.duolingo.com/resources

Thissen, D. & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer.* Mahwah, NJ: Lawrence Erlbaum Publishers.

U.S. Department of Labor. (2006). *Testing and assessment: A guide to good practices for workforce investment professionals.* Retrieved from: https://wdr.doleta.gov/directives/attach/TEN/ten2007/TEN21-07a1.pdf

Wainer, H. (2000). *Computerized adaptive testing: A primer.* New York, NY: Routledge.

Ye, F. (2014). *Validity, reliability, and concordance of the Duolingo English Test.* Retrieved from: https://s3.amazonaws.com/duolingo-papers/other/ye.testcenter14.pdf

Zhu, X., & Goldberg, A. B. (2009). *Introduction of semi-supervised learning*. Synthesis of lectures on artificial intelligence and machine learning. Williston, VT: Morgan & Claypool. Retrieved from https://www.morganclaypool.com/doi/abs/10.2200/S00196ED1V01Y200906AIM006

Back to Table of Contents