Alignment of the Duolingo English Test (DET) to the Common European Framework of Reference for Languages (CEFR): *An External Validation of Individual Skill Subscores for Speaking, Writing, Reading, and Listening*

Kelley Stethen, MEd

Chad Buckendahl, PhD

ACS Ventures, LLC

Submitted to Duolingo: May 2025

Preferred citation:

Stethen, K., & Buckendahl, C. (2025). *Alignment of the Duolingo English Test (DET) to the Common European Framework of Reference for Languages (CEFR): An external validation of individual skill subscores for speaking, writing, reading, and listening*. ACS Ventures, LLC.

1. Introduction

This report presents the results of an external validation conducted by ACS Ventures LLC (ACS) at the request of Duolingo, Inc (Duolingo). The purpose of the study was to gather external validity evidence used to evaluate alignment of the Duolingo English Test (DET) individual skill subscores for speaking, writing, reading, and listening (SWRL) with the *Common European Framework of Reference for Languages* (CEFR) (Council of Europe, 2001, 2020). The report provides a high-level summary of the recommendations resulting from the standards validation study, which will serve as additional evidence in evaluating the alignment of DET cut scores to the standards of proficiency defined by each level of the CEFR.

This section provides high-level background information on the theoretical basis of the CEFR, the unique context of the DET, and the rationale for evaluating score alignment between the two proficiency scales.

1.1. The Common European Framework of Reference (CEFR)

The CEFR was developed by the Council of Europe to support quality in language education by providing standardized reference points for describing language proficiency in support of quality language education (Council of Europe, 2001a, 2020). As a framework, it defines levels of ability along a developmental continuum from pre-A1 to C2, with each level characterized by "can do" statements that describe what language users are able to do in a variety of real-world contexts. These descriptors are organized both globally and by communicative activities which include receptive, productive, interactive, and mediational uses of language and may involve either spoken/signed or written modalities.

The CEFR is intended to enhance transparency in language learning, teaching, and assessment by offering a common framework for comparing results across language assessments (Council of Europe, 2001a, 2020). This comparability supports consistent interpretation and use of test taker performance in high-stakes contexts such as university admissions and employment selection. Given the CEFR's use as a common framework for describing and comparing language proficiency, claims about alignment require justification.

1.2. Evaluating Alignment

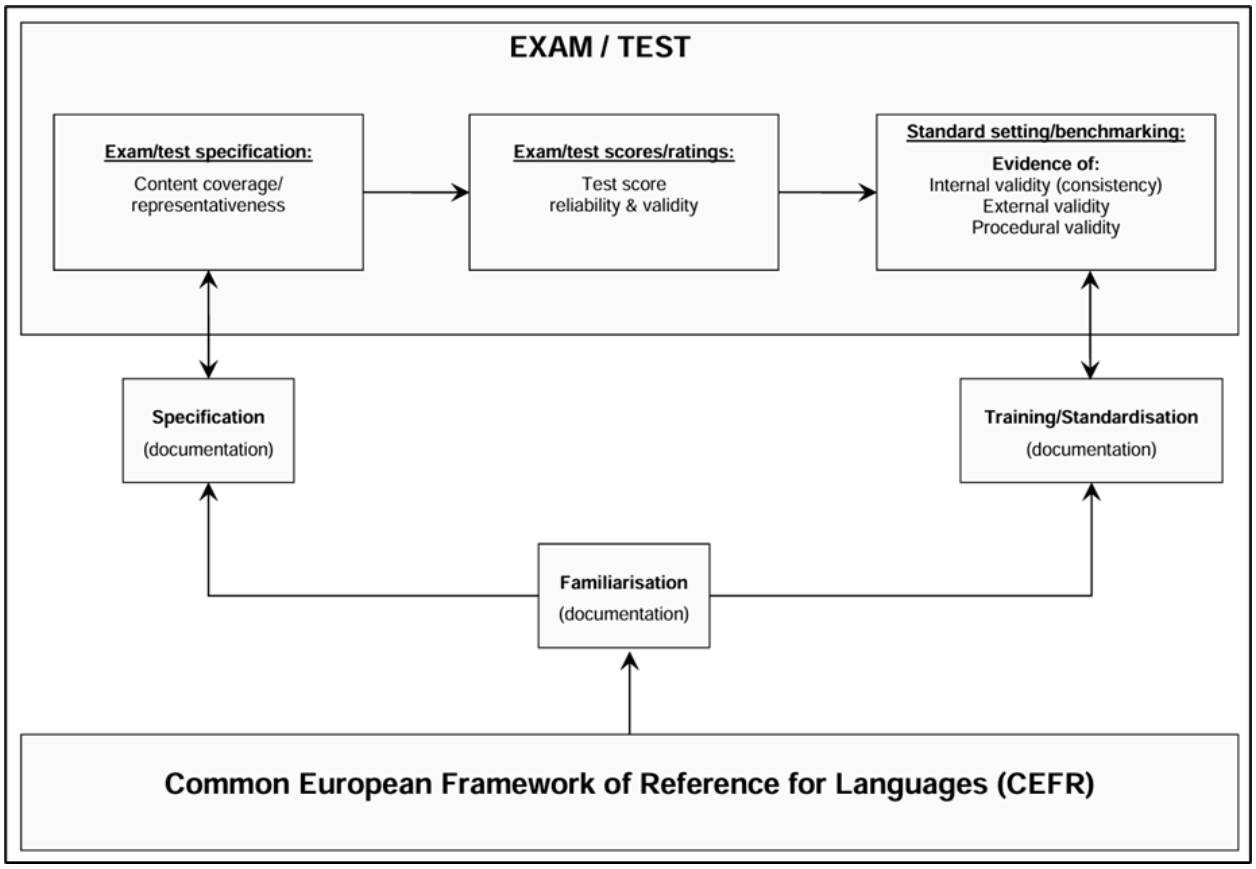
Kane's argument-based approach for validation (2001, 2013) provides a foundation for evaluating claims about test score interpretation and use. This framework emphasizes two core assumptions: (1) that the performance standard is appropriate for the intended score use, and (2) that the cut score accurately reflects that standard. In this context, performance standards represent policy-level expectations for competence, while cut scores serve as operational definitions, identifying the minimum score (i.e., threshold) at which a test taker is considered to meet the expected level. For CEFR-aligned assessments, the CEFR levels function as the

performance standards. Because these levels are predefined, validation efforts focus on evaluating the extent to which test scores align with them.

To guide alignment efforts, the Council of Europe (2009, 2011) published materials outlining procedures for linking (aligning) tests to the CEFR across multiple stages (shown in Figure 1). The procedures resemble Kane's (2001, 2013) framework, and describe alignment as an argument-based process built on claims about specifications (content standards) and standard setting (performance standards), which are corroborated through validation.

Figure 1

Validity Evidence of Alignment of Test Results to the CEFR



Note. From "Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A manual," Council of Europe, 2009, p. 8.

1.3. The Duolingo English Test (DET)

The DET is a computerized adaptive test (CAT) designed to measure English language proficiency across the skills required for effective communication in English-medium settings (Cardwell et al., 2024). The test is delivered online and is available on a continuous basis, allowing test takers to complete the assessment remotely without the need for a physical test center. The DET uses

adaptive algorithms for item selection and employs automated item (task) generation techniques to support the development of item pools without the need for pretesting (LaFlair et al., 2022). The testing model is intended to provide flexible access to English language proficiency certification while maintaining psychometric quality comparable to other standardized English proficiency tests.

DET results are intended for use in high-stakes decision-making contexts, such as including university admissions and employment selection, where verified evidence of English proficiency is required. As with other assessments referencing the CEFR, claims of alignment require evidence that the test is appropriately linked to the levels and descriptors defined by the framework. The following sections present the evidence used to validate claims of alignment based on the specification and standard setting for the DET.

1.3.1. Specification Evidence

The DET assesses test taker ability to use the individual linguistic skills of speaking, writing, reading, and listening (SWRL), as well as the integration of these individual skills for literacy, conversion, comprehension, and production, aligning with the communicative language competencies outlined in the CEFR (Cardwell, et al., 2024). Each linguistic skill is evaluated through a range of task types designed to elicit evidence across the individual and integrated skills. Evidence of content validity and score reliability are summarized below and further documented in the *DET Technical Manual* (Cardwell, et al., 2024).

Content Validity

The DET uses task performance to report an overall score, as well as subscores for each individual SWRL skill (Cardwell, et al., 2024). These subscores are grounded in distinct construct definitions, informed by theoretical frameworks in communicative language ability (e.g., Bachman & Palmer, 1996; Canale & Swain, 1980), and are operationalized through specific task designs. Table 1 provides a high-level summary of the key constructs and task types associated with each individual skill. See the *DET Technical Manual* (Cardwell, et al., 2024) for a more detailed description of the skills and associated tasks.

Table 1

Overview of DET Skills and Task Types

Individual Skill	Description	Integrated Skill	Task Types
Speaking	Producing spoken English from basic discourse to advanced discourse at CEFR levels A1-C2	Conversion; Production	● Extended Speaking (audio prompt; text prompt) ● Picture Description (speaking) ● Read Aloud ● Speaking Sample
Writing	Writing English texts with a variety of rhetorical functions at CEFR levels A1-C2	Literacy; Production	● Interactive Writing ● Interactive Listening (S) ● Picture Description (writing) ● Writing Sample
Reading	Comprehending written English from basic informational texts to advanced expository/persuasive texts at CEFR levels A1-C2	Literacy; Comprehension	● Interactive Reading ● C-test ● Vocabulary in Context ● Yes/No Vocabulary
Listening	Comprehending spoken English from basic discourse to advanced discourse at CEFR levels A1-C2	Conversation; Comprehension	● Interactive Listening (DC) ● Interactive Listening (S) ● Extended Speaking (audio prompt) ● Dictation

Note. From "Duolingo English Test: Technical Manual," Cardwell et al., 2024, p. 5.
S = summarization, DC = dialogue completion.

Score Reliability

The DET's scoring system is designed to ensure consistency across test forms and test takers through a combination of machine scoring and human-rated responses. Reliability estimates for the overall score and subscores from the individual SWRL skills are reported using appropriate psychometric methods (e.g., marginal reliability, test-retest correlations), with coefficients typically exceeding accepted thresholds for high-stakes assessments. The scoring model accounts for shared variance across tasks to provide a more holistic view of communicative competence, consistent with the multidimensional nature of language use described in

frameworks like the CEFR. Additionally, the adaptive nature of the test further contributes to measurement precision by targeting item difficulty to each test taker's proficiency level.

1.3.2. Standard Setting Evidence

Standard setting was conducted in the form of an empirical study designed to examine the concordance between DET scores and scores from other similar English language proficiency tests (Cardwell, et al., 2023). The standard setting study employed a known-groups design and kernel equating to compare self-reported and verified DET scores on the overall score to performance on other CEFR-aligned English language assessments.

The results of this study provided initial evidence of alignment with the CEFR and were used to establish preliminary cut scores on the DET overall scale for each CEFR level (shown in Table 2 below).

Table 2

Cut Score Correspondence

CEFR Level	Corresponding Cut Score
C2	155
C1	130
B2	100
B1	60
A1-A2	10-55

Note. A1-A2 combined for stability

Validating claims based on standard setting validity requires evaluating evidence that the resulting cut scores align with the performance standard defined by the CEFR levels (Kane, 2001, 2013). These claims can be evaluated by reviewing standard setting evidence documenting the internal consistency of the data generated (*internal validity*), the appropriateness and fidelity of the procedures used (*procedural validity*), and external evidence supporting the validity of using the recommended cut scores (*external validity*) (Council of Europe, 2009; Kane, 2001, 2006). The evaluation of the standard setting validity is presented below, organized by these sources of evidence.

Internal Validity

Internal validity is concerned with the accuracy and consistency of the standard setting results (Council of Europe, 2009; Kane, 2001, 2006). In the standard setting study by Cardwell et al. (2023), internal validity was supported through a sensitivity analysis designed to assess the consistency of the results. This approach involves repeating the equating procedure across

multiple conditions and comparing the resulting score concordances. This analysis also included an examination of conditional standard errors of equating (SEE) to evaluate the accuracy of the equating process by quantifying the amount of error that is introduced (see Cardwell et al. [2023] Online Supplemental materials for technical details on the method). Across all tested conditions, the selected concordance method produced results closely aligned with the median, providing evidence in support of the internal validity of the cut scores.

Procedural Validity

Procedural validity is concerned with the sufficiency of the standard setting procedures (Council of Europe, 2009; Kane, 2001, 2006). The approach used by Cardwell et al. (2023) identifies score ranges that minimize misclassification error and is conceptually similar to *Contrasting Groups* or *Borderline Group* methods commonly used in standard settings (Hambleton & Pitoniak, 2006). This study was grounded in empirical data to produce cut scores on the overall scale that are aligned to the threshold levels of the CEFR. However, the *Standards for Educational and Psychological Testing* (AERA et al., 2014) emphasize the importance of incorporating content-based judgments when evaluating the validity of standard setting procedures. Therefore, additional evidence of validity from other standard setting procedures would be needed to support the claim of alignment.

External Validity

External validation is a process that can be used to gather evidence from additional independent sources to corroborate the results of a standard setting (Council of Europe, 2009; Kane, 2001, 2006). One way to collect this evidence involves referencing the results of a standard setting study to an external criterion that is known to be reliably calibrated as a test of the same skill. Another way is through cross validation (referred to here as *standards validation*), where multiple standard setting studies are conducted using the same panel or an independent panel and comparing the results.

A standards validation study was conducted by ACS to evaluate the alignment of the DET overall score scale with the CEFR levels using content-based judgments from subject-matter experts (Stethen, et al., 2024). The study followed established standard setting procedures and included a virtual workshop in which trained panelists reviewed CEFR descriptors, DET test specifications, and representative test taker performance presented in an Ordered Item Booklet (OIB). The OIB approach integrated features of the *Bookmark*, *Item Descriptor Matching*, and *Body of Work* methods to support holistic judgments of threshold performance based on test content (Ferrera, et al., 2008; Kingston et al., 2012).

In the standards validation study by Stethen et al. (2024), panelists submitted independent cut score recommendations for CEFR levels, followed by group discussion and final consensus

ratings. As shown in Table 3 below, panelist recommendations were closely aligned with existing DET cut scores from a prior empirical standard setting study, with the empirical cuts falling into the recommended ranges for levels B2, C1, and C2, and suggesting a minor upward revision for the B1 threshold. The low standard error of the medians[1] ($SE_{Medians}$) and narrow confidence intervals indicated strong agreement among panelists, supporting the validity of the DET's overall cut score alignment to the levels of the CEFR.

Table 3

Overall Cut Score Recommendations

Recommendation	**Overall Cut Score**		
Level *(Cut Score)*	***Median***	***SE_{Median}***	***95% CI***
C2 *(155)*	150	0.0	150 - 155
C1 *(130)*	130	0.8	130 - 135
B2 *(100)*	100	0.0	100 - 100
B1 *(60)*	65	0.0	65 - 65

While the previous standard setting and validation efforts have established alignment between the DET overall score and CEFR proficiency levels, the DET also reports subscores for the individual SWRL skills. As these subscores are interpreted and used alongside the overall score, it is critical to ensure that each subscore independently aligns with the CEFR levels with comparable validity. To address this, an additional standards validation study was conducted by ACS to evaluate the alignment of the DET's individual skill SWRL subscores with the CEFR levels. This study applied a parallel methodology to the prior standards validation of the overall score, using content-based judgments from trained subject-matter experts (SMEs). Together, the two studies provide external validity evidence supporting the interpretation and use of both the DET's overall and individual skill subscores in relation to CEFR-aligned performance standards.

2. Methods

To evaluate the validity of score alignment between the Duolingo English Test (DET) and the Common European Framework of Reference for Languages (CEFR), the DET Assessment Research team contracted with ACS Ventures, LLC (ACS) to conduct an external validation of the empirical cut scores. This occurred in the form of two standards validation studies designed to examine whether the cut scores from the previous standard setting appropriately reflect the performance expectations defined by each CEFR level for the overall score and SWRL subscores for the individual skills.

[1] SE_{Median} for the overall score recommendations was estimated by multiplying the standard error of the mean by a constant value of 1.2533 (Rider, 2012).

The following section of this report describes the methodology used in the standards validation study of the individual skill subscores, including procedures for panel selection, task materials, rating processes, and analytical strategies used to evaluate alignment outcomes.

2.1. Participants

Thirteen participated in the standards validation study for each individual SWRL skill. The final panelists were selected through efforts by DET Assessment Research, and input from ACS and Dr. Anthony Green, a language testing consultant with expertise in English language tests, with the intent to ensure there was a diversity of nationalities, geographic locations, linguistic backgrounds (including English varieties), and work experiences. Additional efforts were made to select panelists who were familiar with the target testing population and who were subject matter experts (SMEs) in Teaching English as a Second or Other Language (TESOL) content, but all panelists were provided with training/retraining on the CEFR and the specifications of the DET as a part of the pre-workshop familiarization activities.

Panelist demographic information was collected as part of a pre-workshop survey. This survey collected self-reported information about nationality, first language, geographic location, gender, educational and professional background, as well as teaching and assessment experience. A summary of the demographics from the panel is presented in Table 4 below. The complete set of information can be found in Appendix A.

Table 4

SWRL Subscore Study (N = 13)

Demographic Information	n	%
Education		
Bachelor's degree	0	0%
Master's degree	8	62%
Doctoral degree	5	38%
First Language [a]		
Chinese	1	8%
English	7	54%
Mandarin/Cantonese	1	8%
Hebrew	1	8%
Russian	2	15%
Spanish	2	15%
Teaching Experience		
Less than 5 years	0	0%
5 to 10 years	2	15%
11 to 20 years	4	31%
20 or more years	7	54%
Assessment Experience		
Less than 5 years	1	8%
5 to 10 years	3	23%
11 to 20 years	7	54%
20 or more years	2	15%

Note. [a] Multiple First Languages could be reported

Table 4 shows the panel for the SWRL study had representation from a range of first languages and advanced levels of education (degrees held by the panelists were typically in the fields of education, applied linguistics, TESOL, etc.). Overall, the panelists ranged from 9 to 27 years of teaching experience and an average of 13.4 years of experience in assessment.

2.2. Materials

All materials used were designed to support expert panelists in evaluating the alignment of DET scores to the CEFR levels for the standards validation study. Materials included test specifications, CEFR performance descriptors, sample DET items, exemplars of test taker performance, and rating forms tailored to the standard-setting procedures.

2.2.1. Familiarization

Prior to the start of the workshop, panelists were required to complete a series of pre-workshop activities that were designed to familiarize themselves with the CEFR and the DET, including expectations associated with each level of the CEFR. These activities were selected by Duolingo to reflect the recommendations provided in the Manual and were reviewed by Dr. Green, who provided expertise on which illustrative examples would be the most appropriate for the task. Panelists were expected to spend approximately five hours reviewing these pre-workshop activities in preparation for the workshop.

A summary description of the materials provided for familiarization activities is provided below and a more detailed description of each is included in the *Supplemental Materials*.

CEFR Descriptors and Illustrative Examples

To support the evaluation of CEFR alignment, panelists received a curated set of CEFR descriptors, including both global proficiency level summaries and skill-specific "can-do" statements drawn from the CEFR Companion Volume (Council of Europe, 2020). These descriptors functioned as the performance standards against which panelists evaluated the appropriateness of DET score thresholds. As part of their pre-workshop familiarization, panelists engaged in structured activities that required them to review and apply CEFR descriptors in context. These activities emphasized the expectations associated with each CEFR level and guided panelists in interpreting the meaning and implications of the descriptors across the SWRL skills. Developed in accordance with Council of Europe recommendations (2009, 2011, 2020), these familiarization exercises were reviewed by Dr. Anthony Green to ensure they reflected appropriate and illustrative examples. The goal was to enhance consistency and confidence in panelists' use of the CEFR during the standard-setting process.

DET Specifications and Sample Tasks

Panelists were also provided with detailed documentation outlining the structure of the DET, including descriptions of the test's adaptive algorithm, task types, scoring model, and reported score scales. These familiarization materials included targeted activities designed to help them understand the constructs measured by each task type and how these constructs align with communicative language proficiency. The goal of these familiarization activities was to ensure that panelists could make judgments with an informed understanding of how the DET tasks operationalize language constructs relevant to the CEFR. Panelists were expected to dedicate approximately five hours to this familiarization process.

2.2.2. Ordered Item Booklets (OIBs)

Evaluating the cut score validity for a CAT, such as the DET, requires representation of the diversity of features and characteristics of the full item bank. To best accommodate this CAT

design feature, a methodology that allowed the use of an Ordered Item Booklet (OIB) to represent the range of content, task types, and test taker performance across the score scale was determined to be the most appropriate. The methodology integrates elements from established standard-setting procedures such as the Bookmark and Item Descriptor (ID) Matching methods (Ferrara, et al. 2008), while also incorporating elements from the Body of Work method (Kingston et al., 2012). By synthesizing these approaches, the OIB methodology supported judgments grounded in both task features and integrated language performance.

The same broad OIB methodology from the standards validation of the overall score by Stethen at al. (2024) was applied in this standards validation of the individual skill subscores. The main difference between the OIBs for the overall score study and the subscore study is that performance needed to be evaluated in detail for each of the subscores. Therefore, separate OIBs were created and recommendations made for each individual SWRL skill.

Detailed criteria for the construction and use of the OIBs are provided in the *Supplemental Materials*.

2.2.3. Rating Forms and Procedural Guides

Panelists received a set of structured materials designed to support the standard-validation process and guide their judgments regarding score alignment. These included online rating forms for recording cut score recommendations and comments, and procedural instructions for how to use these materials to provide a recommended cut score for each CEFR level. To facilitate the task, panelists were provided with: (a) a summary of how the original DET cut scores were determined; (b) descriptors of CEFR-level performance; (c) information on task selection and scoring procedures; and (d) the existing DET cut score locations on the relevant score scale. These materials were referenced throughout the training and rating process to ensure consistency in interpretation and use.

Panelists submitted their recommendations through an online form (see *Supplemental Materials* for an example form), which was used to collect individual cut score decisions prior to group discussion and consensus-building. The following section outlines the procedures used in this SWRL standards validation workshop.

2.3. Procedures

The standards validation study was conducted as a virtual workshop using online meeting software. To accommodate panelists across multiple time zones, the workshops were delivered in a hybrid format, combining synchronous meetings with offline, independent review activities. The following sections describe the procedures, including pre-workshop activities, panel orientation and training, and the recommendation processes.

2.3.1. Pre-Workshop Activities

Prior to the start of the workshop, panelists were required to complete a set of familiarization activities designed to build foundational knowledge of both the CEFR and the DET. These activities focused on understanding the expectations associated with each CEFR level and the constructs measured by DET tasks. Materials were developed in accordance with guidance from the Council of Europe (2009, 2011, 2020) and reviewed by Dr. Anthony Green to ensure the appropriateness and representativeness of illustrative examples. Panelists were expected to spend approximately five hours completing the activities, which included reviewing CEFR descriptors, DET task descriptions, and sample test-taker responses.

The full set of familiarization activities is provided in the *Supplemental Materials*.

2.3.2. Orientation and Study Overview

The workshop began with a live general orientation session that introduced the panelists to the study's goals, design, and procedures. The facilitation team from ACS provided an overview of the purpose of the validation study, describing how the current results would be used alongside prior standard setting and standards validation efforts to evaluate the validity of score alignment claims to CEFR levels. Dr. Ben Naismith from Duolingo delivered a presentation on the DET's intended use, scoring model, and score interpretation frameworks.

2.3.3. Standardization and Training

Following the orientation, panelists engaged in structured training activities. First, they reviewed performance expectations across CEFR levels using the CEFR descriptors for the relevant skills. Panelists were then asked to reflect on the key knowledge and skills that differentiate performance at the threshold between CEFR levels. This review was followed by DET-specific training, which included discussion of task types, scoring models, and the constructs assessed by each task.

Panelists were introduced to the OIBs, which provide a representation of test taker performance from across the DET scale. The SWRL study required panelists to use four separate OIBs, each demonstrating performance associated with a specific SWRL skill. The performance examples provided in the OIBs were designed to support holistic evaluation of candidate ability at and around each CEFR threshold.

To prepare panelists for the recommendation task, the facilitator provided specific instruction on how to use the OIBs to identify whether the existing cut scores (established in prior studies) accurately reflected CEFR threshold performance. Training included practice rounds using the

B1 and C1 thresholds as examples. Panelists were also provided with the following reference materials to support their judgments:

- A summary of how original DET cut scores were determined
- CEFR descriptors for the relevant skills
- Information on task selection and scoring
- The current DET score scale and cut score locations

2.3.4. Providing Recommendations

Panelists began by submitting individual *initial recommendations* using an online form (see *Supplemental Materials*). These recommendations identified the score point (which corresponded to the tab of a spreadsheet document) that best reflected the threshold performance for each CEFR level. Panelists also rated their confidence in their recommendations.

Once initial recommendations were submitted, the panel reconvened to discuss their evaluations. Panelists shared their rationale, noted areas of uncertainty, and discussed how they interpreted candidate performance within the OIBs. The facilitator presented summary statistics (e.g., mean, median, and distribution of cut score selections), which served as a basis for further review and discussion. Panelists were asked to re-examine the score points surrounding each CEFR threshold, considering the group discussion and their own review of the OIBs.

Using this process, panelists refined their recommendations and submitted *final recommendations* regarding whether the existing cut scores were appropriate or should be adjusted. Final recommendations were based on the median of panelist judgments. In both studies, this approach provided a consensus estimate of the score point that best represented CEFR threshold expectations.

A summary of the final recommendations is presented in the following section of this report.

3. Results

This section presents the results of the standards validation study evaluating the alignment of Duolingo English Test (DET) individual skill subscores for the skills of speaking, writing, reading, and listening (SWRL) with the *Common European Framework of Reference for Languages* (CEFR). Results include panelist recommendations, confidence ratings, and statistical indicators of consensus, as well as an evaluation of the procedural, internal, and external validity evidence.

3.1. Panel Recommendations

ACS compiled the recommended cut scores, along with the panelists' confidence in the recommendations, and calculated the median, standard error of the median[2] (SE_{Median}), and 95% confidence intervals (bounded by 5-point intervals) for each level of the CEFR.

Panelists evaluated cut scores for each individual SWRL skill. Initial and final recommendations, along with confidence intervals and standard error of the median, are presented in Tables 5–8 (see Appendix B for a summary of recommendations).

Table 5

Subscore Recommendations: Speaking Skill

Overall Scale		**Initial**			**Final**	
Level *(Cut Score)*	***Median***	***SE_{Median}***	***95% CI***	***Median***	***SE_{Median}***	***95% CI***
C2 (155)	150	2.4	150 - 155	150	2.4	150 - 155
C1 (130)	130	2.5	130 - 135	130	2.4	130 - 135
B2 (100)	100	0.5	100 - 100	100	0.0	100 - 100
B1 (60)	65	0.0	65 - 65	65	0.0	65 - 65

Table 6

Subscore Recommendations: Writing Skill

Overall Scale		**Initial**			**Final**	
Level *(Cut Score)*	***Median***	***SE_{Median}***	***95% CI***	***Median***	***SE_{Median}***	***95% CI***
C2 (155)	150	1.7	148 - 155	150	0.0	150 - 150
C1 (130)	130	0.9	130 - 130	130	0.7	130 - 130
B2 (100)	100	1.5	95 - 100	100	2.5	95 - 100
B1 (60)	60	1.9	60 - 65	60	1.9	60 - 65

Table 7

Subscore Recommendations: Reading Skill

Overall Scale		**Initial**			**Final**	
Level *(Cut Score)*	***Median***	***SE_{Median}***	***95% CI***	***Median***	***SE_{Median}***	***95% CI***
C2 (155)	150	2.5	150 - 155	155	2.0	150 - 155
C1 (130)	130	0.5	130 - 130	130	0.0	130 - 130
B2 (100)	95	1.9	95 - 100	95	2.5	95 - 100
B1 (60)	60	2.0	55 - 60	60	1.3	55 - 60

[2] For the subscore recommendations, the SE_{Median} was estimated using a bootstrap resampling approach (Efron, 1979).

Table 8

Subscore Recommendations: Listening Skill

Overall Scale	**Initial**			**Final**		
Level *(Cut Score)*	***Median***	***SE_{Median}***	***95% CI***	***Median***	***SE_{Median}***	***95% CI***
C2 (155)	155	0.0	155 - 155	155	0.6	155 - 155
C1 (130)	130	0.0	130 - 130	130	0.0	130 - 130
B2 (100)	100	1.3	100 - 105	100	0.6	100 - 100
B1 (60)	60	0.2	60 - 60	60	0.2	60 - 60

Panelists' recommendations generally supported the existing DET cut scores, though several differences were noted:

- ***Speaking***: Panelists recommended raising the B1 cut score to 65. There was also discussion of lowering the C2 cut score to 150; however, there was not full consensus, resulting in a recommended range of 150–155.
- ***Writing***: Panelists unanimously recommended lowering the C2 cut score from 155 to 150.
- ***Reading***: While most panelists recommended retaining current cut scores, several recommended lowering the B2 cut score to 95, resulting in a suggested range of 95–100.
- ***Listening***: Final recommendations were fully aligned with the existing overall DET cut scores.

These differences provide nuanced evidence for considering minor adjustments at specific skill levels but affirm the general alignment of DET subscore thresholds with CEFR expectations.

3.2. Evaluation of Validity Evidence

3.2.1. Procedural Validity

Procedural validity evidence encompasses the appropriateness of participant selections, the adequacy of methods used, and the participant confidence in the process (Kane, 2001, 2006). Panelists were carefully selected based on their expertise and familiarity with the CEFR framework and language teaching experience, ensuring the appropriateness of their judgments. Comprehensive training sessions helped strengthen understanding of the judgment task and processes.

All panelists reported feeling confident in their ability to make informed recommendations following the training and practice rounds. Survey data showed increases in panelist confidence from the initial to final rounds of the study (see Table 9 below).

Table 9

Confidence in Recommendations

Recommendation Confidence		Not at all Confident		Not Very Confident		Somewhat Confident		Very Confident	
		n	*%*	*n*	*%*	*n*	*%*	*n*	*%*
Speaking									
	Initial (N=13)	0	0%	1	8%	9	69%	3	23%
	Final (N = 13)	0	0%	0	0%	5	38%	8	62%
Writing									
	Initial (N=13)	1	8%	0	0%	8	62%	4	31%
	Final (N = 13)	0	0%	0	0%	6	46%	7	54%
Reading									
	Initial (N=13)	0	0%	2	15%	11	85%	0	0%
	Final (N = 13)	0	0%	0	0%	9	69%	4	31%
Listening									
	Initial (N=13)	0	0%	6	46%	7	54%	0	0%
	Final (N = 13)	0	0%	0	0%	9	69%	4	31%

3.2.2. Internal Validity

Internal validity evidence evaluates the consistency and reliability of the panelists' recommendations (Kane, 2001, 2006). A key finding is the trend in the (SE_{Median}) reducing from the initial to final rounds of recommendations.

Table 10 presents the trend in the change (Δ) in the SE_{Median} between rounds for each cut score recommendation.

Table 10

Change in SE_{Median} between initial and final recommendations

Cut Score SE_{Median}	***Δ SE_{Median}***			
	B1	***B2***	***C1***	***C2***
Speaking	0.0	(-0.5)	(-0.1)	(-0.2)
Writing	0.0	(-1.0)	(-0.2)	(-1.7)
Reading	(-0.7)	0.6	(-0.5)	(-0.5)
Listening	0.0	(-0.7)	0.0	0.6

This trend indicates an increase in panelist agreement, demonstrating improved reliability of the cut score recommendations as the process progressed.

3.2.3. External Validity

External validity evidence considers whether the study outcomes align with prior research or independent validation sources (Kane, 2001, 2006). This study drew on prior standard setting and standard validation studies that established the alignment of the DET scores with the levels of the CEFR.

Table 11 provides a comparison of the final recommendations compared to the overall cut scores resulting from the previous standard setting and standards validation study (Cardwell, et al., 2023).

Table 11

Comparison of Subscore Cut Score Recommendations to Existing Overall Cut Scores

Score Scale	CEFR Level *(Cut Score)*			
	B1 ***(60)***	**B2** ***(100)***	**C1** ***(130)***	**C2** ***(155)***
Overall				
Median (SE_{Median})	65 (0.0)*	100 (0.0)	130 (2.4)	150 (2.4)*
95%CI_{Median}	65 - 65	100 - 100	130 - 135	150 - 155
Speaking				
Median (SE_{Median})	65 (0.0)*	100 (0.0)	130 (2.4)	150 (2.4)*
95%CI_{Median}	65 - 65	100 - 100	130 - 135	150 - 155
Writing				
Median (SE_{Median})	60 (1.9)	100 (2.5)	130 (0.7)	150 (0.0)*
95%CI_{Median}	60 - 65	95 - 100	130 - 130	150 - 150
Reading				
Median (SE_{Median})	60 (1.3)	95 (2.5)*	130 (0.0)	155 (2.0)
95%CI_{Median}	55 - 60	95 - 100	130 - 130	150 - 155
Listening				
Median (SE_{Median})	60 (0.2)	100 (0.6)	130 (0.0)	155 (0.6)
95%CI_{Median}	60 - 60	100 - 100	130 - 130	155 - 155

Note. (*) indicates recommended change from overall cut score

The findings from this subscore standards-validation study reaffirmed the conclusions of the previous studies, demonstrating that the recommended SWRL cut scores are generally consistent with the conceptual expectations of performance described by the CEFR levels.

3.3. Summary of Findings

At the conclusion of the study, the cut score recommendations for the thresholds of the CEFR levels were provided to Duolingo for their review. Based on the evidence collected and panelists' performance, the results of the standards validation study generally supported the current DET cut scores, while also identifying specific areas where refinements could be considered for the subscores associated with the individual SWRL skills (see Table 11).

Using Kane's (2001, 2006) framework to evaluate procedural, internal, and external validity evidence, this study offers strong support for the claim that the DET aligns with the CEFR levels at the subscore scale. These evaluation results support the use of the DET as a measure of language proficiency and affirm its appropriateness for use in high-stakes decision-making contexts. ACS recommends the DET Assessment Research Team consider these recommendations, along with other evidence such as qualitative feedback from panelists and the results of other external studies, to determine the final cut scores for each subscore scale.

References

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

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

Canale, M., & Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. Applied Linguistics, 1, 1-47. http://dx.doi.org/10.1093/applin/I.1.1

Cardwell, R. L., Naismith, B., LaFlair, G. T., & Nydick, S. (2024). Duolingo English Test: Technical Manual. *Duolingo Research Report*. https://go.duolingo.com/dettechnicalmanual

Cardwell, R. L., Nydick, S. W., Lockwood, J. R., & von Davier, A. A. (2023). Practical considerations when building concordances between English tests. *Language Testing*, 41(1), 192-202. https://doi.org/10.1177/02655322231195027

Council of Europe (2001a): Common European Framework of Reference for Languages: learning, teaching, assessment. Cambridge: Cambridge University Press.

Council of Europe (2009). Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). Council of Europe. https://www.coe.int/en/web/common-european-framework-reference-languages/relating-examinations-to-the-cefr

Council of Europe (2011). *Manual for Language Test Development and Examining*: *For use with the CEFR*. Council of Europe.

Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7(1), 1-26. https://doi.org/10.1214/aos/1176344552

Ferrara, S., Perie, M., & Eugene, J. (2008). Matching the Judgmental Task with Standard Setting Panelist Expertise: The Item-Descriptor (ID) Matching Method. *Journal of Applied Testing Technology*, 9(1), 1-22.

Hambleton, R. K. & Pitoniak, M. J. (2006): Setting Performance Standards. In Brennan, R.L. (Ed.) Educational Measurement (4th ed., pp. 433-470). Westport, CT: American Council on Education/Praeger.

Kane, M. (2001). Current concerns in validity theory. Journal of Educational Measurement, *38*(4), 319-342.

Kane, M. (2006). Validation. In R. B. Brennan (Ed.), Educational Measurement (4th ed., pp. 17-64). American Council on Education/Praeger.

Kane, M. (2013). Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement, 50(1), 1-73.

Kingston, N. M. & Tiemann, G. C. (2012). In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, methods, and innovations* (2nd ed., pp. 194-216). Routledge.

Mitzel, H. C., Lewis, D. M., Patz, R. J. & Green, D. R. (2001): The Bookmark Procedure: psychological perspectives. In Cizek G. J. (Ed.) Setting performance standards: Concepts, methods, and perspectives (pp. 249-281). Mahwah, NJ: Erlbaum.

Rider, P. R. (1960). Variance of the Median of Small Samples from Several Special Populations. *Journal of the American Statistical Association*, *55*(289), 148–150. https://doi.org/10.1080/01621459.1960.10482056

Stethen, K., Buckendahl, C., Green, A. (2024). *Alignment of the Duolingo English Test (DET) to the Common European Framework of Reference for Languages (CEFR): An External Validation* [Unpublished manuscript]. Duolingo English Test Assessment Research.

Appendices

Appendix A: Panelist Information

Table A. 1. SWRL Study Panelist Background Information

ID	Nationality	L1(s)	Gender	Degree	Role	Teaching Experience (years)	Assessment Experience (years)
P1	US	English	Male	MA TESOL	Language assessment specialist	9	20
P2	Canadian	English	Female	BA Linguistics, MEd (TESL)	Literacy instructor	25	15
P3	US	English	Female	PhD Second Language Education	Language Teacher	12	3
P4	US	Russian	Female	PhD Second Language Acquisition and Teaching	Project coordinator/Item reviewer	15	13
P5	Canadian	English	Female	MEd	Professor	25	18
P6	Israeli	Hebrew	Male	MEd	Teacher/Trainer/ Materials writer	20	15
P7	UK/ Norwegian	English	Male	MPhil Second Language Education	PhD researcher (Teacher Education)	12	8
P8	Mexican	Spanish	Female	PhD Applied Linguistics	Professor, Head of a Foreign Languages Department	27	17
P9	Canada/ Hong Kong	Mandarin/ Cantonese/ English	Male	MA Applied Linguistics	Teacher Trainer/PhD candidate	20	15
P10	British/ Canadian	English	Male	MA TESOL, DELTA	Teacher, Trainer, & Assessor	25	20
P11	Canadian	Spanish	Male	PhD Candidate; MEd	Manager of Language Services	20	10
P12	Russian	Russian	Male	PhD Applied Linguistics	Associate Professor of English/TESOL	15	7
P13	Chinese	Chinese	Female	PhD	Assessment Lead at a University	10	13

Appendix B: Summary of Recommendations

Table B. 1. Speaking Recommendations

Summary of Panelist Ratings: Speaking				
Initial Recommendations				
Score by CEFR Level	***B1***	***B2***	***C1***	***C2***
N Responses	13	13	13	13
Min	60	100	125	145
Max	65	110	135	155
Mean	64.6	101.5	131.9	151.9
SD	1.39	3.15	3.25	3.25
SE	0.38	0.87	0.90	0.90
95% CI [upper] Mean	65	103	134	154
95% CI [lower] Mean	64	100	130	150
Median	65	100	130	150
SEM	0.00	0.52	2.45	2.42
95% CI [upper] Median	65	100	135	155
95% CI [lower] Median	65	100	130	150
Final Recommendations				
Score by CEFR Level	***B1***	***B2***	***C1***	***C2***
N Responses	13	13	13	13
Min	60	95	125	145
Max	65	105	135	155
Mean	64.6	100.4	131.9	151.9
SD	1.39	2.47	3.25	3.25
SE	0.38	0.68	0.90	0.90
95% CI [upper] Mean	65	102	134	154
95% CI [lower] Mean	64	99	130	150
Median	65	100	130	150
SEM	0.00	0.00	2.44	2.44
95% CI [upper] Median	65	100	135	155
95% CI [lower] Median	65	100	130	150

Table B. 2. Writing Recommendations

Summary of Panelist Ratings: Writing				
Initial Recommendations				
Score by CEFR Level	***B1***	***B2***	***C1***	***C2***
N Responses	13	13	13	13
Min	55	95	130	145
Max	65	105	135	155
Mean	61.5	99.6	131.2	150.4
SD	3.15	3.80	2.19	3.96
SE	0.87	1.05	0.61	1.10
95% CI [upper] Mean	63	102	132	153
95% CI [lower] Mean	60	97	130	148
Median	60	100	130	150
SEM	1.91	1.51	0.78	1.73
95% CI [upper] Median	65	100	130	155
95% CI [lower] Median	60	95	130	148
Final Recommendations				
Score by CEFR Level	***B1***	***B2***	***C1***	***C2***
N Responses	13	13	13	13
Min	60	95	130	145
Max	65	105	135	155
Mean	61.9	98.5	131.2	150.0
SD	2.53	3.76	2.19	2.89
SE	0.70	1.04	0.61	0.80
95% CI [upper] Mean	63	101	132	152
95% CI [lower] Mean	60	96	130	148
Median	60	100	130	150
SEM	1.91	2.48	0.65	0.00
95% CI [upper] Median	65	100	130	150
95% CI [lower] Median	60	95	130	150

Table B. 3. Reading Recommendations

Summary of Panelist Ratings: Reading				
Initial Recommendations				
Score by CEFR Level	***B1***	***B2***	***C1***	***C2***
N Responses	13	13	13	13
Min	50	85	120	145
Max	65	105	130	155
Mean	58.5	96.5	128.5	151.5
SD	4.27	4.74	3.15	3.76
SE	1.19	1.31	0.87	1.04
95% CI [upper] Mean	61	99	130	154
95% CI [lower] Mean	56	94	127	149
Median	60	95	130	150
SEM	1.95	1.91	0.52	2.47
95% CI [upper] Median	60	100	130	155
95% CI [lower] Median	55	95	130	150
Final Recommendations				
Score by CEFR Level	***B1***	***B2***	***C1***	***C2***
N Responses	13	13	13	13
Min	50	95	125	145
Max	65	105	130	155
Mean	58.5	97.7	129.6	152.7
SD	3.76	3.30	1.39	3.30
SE	1.04	0.92	0.38	0.92
95% CI [upper] Mean	61	100	130	155
95% CI [lower] Mean	56	96	129	151
Median	60	95	130	155
SEM	1.26	2.45	0.00	2.04
95% CI [upper] Median	60	100	130	155
95% CI [lower] Median	55	95	130	150

Table B. 4. Listening Recommendations

Summary of Panelist Ratings: Listening				
Initial Recommendations				
Score by CEFR Level	***B1***	***B2***	***C1***	***C2***
N Responses	13	13	13	13
Min	55	90	125	145
Max	70	105	135	155
Mean	60.8	100.8	130.0	153.5
SD	3.44	4.00	2.04	3.76
SE	0.96	1.11	0.57	1.04
95% CI [upper] Mean	63	103	131	156
95% CI [lower] Mean	59	98	129	151
Median	60	100	130	155
SEM	0.16	1.29	0.00	0.00
95% CI [upper] Median	60	105	130	155
95% CI [lower] Median	60	100	130	155
Final Recommendations				
Score by CEFR Level	***B1***	***B2***	***C1***	***C2***
N Responses	13	13	13	13
Min	60	100	130	145
Max	70	105	130	155
Mean	61.2	101.2	130.0	153.5
SD	3.00	2.19	0.00	3.76
SE	0.83	0.61	0.00	1.04
95% CI [upper] Mean	63	102	130	156
95% CI [lower] Mean	59	100	130	151
Median	60	100	130	155
SEM	0.22	0.59	0.00	0.55
95% CI [upper] Median	60	100	130	155
95% CI [lower] Median	60	100	130	155