BRITISH COUNCIL

EnglishScore

Speaking Test Purpose and Content

EnglishScore
Produced together with the Centre For Research In English Language Learning And Assessment (CRELLA) at the University of Bedfordshire.

May 2023

Table of contents

1. The purpose and use of EnglishScore

I. The purpose and use of EnglishScore

Who should take EnglishScore?

EnglishScore is an international assessment taken by young adult (16 and 17) and adult (18 and over) learners of English worldwide. Test-takers may come from any language background and any region of the world.

The meaning of EnglishScore results

EnglishScore provides evidence of proficiency in understanding and using English in everyday life and the workplace.

The test is primarily concerned with the *occupational*, *public* and *personal* domains[1] with items that are more *personal* at the lowest levels of difficulty but that focus more on the *public* and then the *occupational* domains as the difficulty level increases.

It relates to a wide range of contexts of language use[2] with a focus on common workplace and social contexts.

The impact of using EnglishScore

EnglishScore aims to encourage people around the world to unlock the potential of the English language by certifying their scores, helping them to prove their level to potential employers. For universities, employers and other organisations, EnglishScore provides a cost-effective means of large-scale English language testing that is used to inform professional development initiatives, course placements and recruitment efforts.

Ownership of EnglishScore

EnglishScore is owned and administered as a joint venture between the British Council (www.britishcouncil.org), the United Kingdom's international organisation for cultural relations and educational opportunities, and Blenheim Chalcot (www.blenheimchalcot.com), a digital venture builder.

Use of EnglishScore results

EnglishScore can be used by employers, universities and governments to assess a test-taker's general English proficiency. Results can be used by employers to benchmark the proficiency of

[1] See Council of Europe (2001, pp.10, 14, 42–100) for information on domains.

[2] See Council of Europe (2001, pp.30, 101–130) for more information on communicative language competences.

their workforce or assess a future employee's language ability, by universities and language schools as a placement or progress measure for their students or by governments and other stakeholders as an index of a learner’s general English proficiency. In addition, English language learners themselves can use the test to understand their level in relation to the CEFR, set individual language learning goals and select appropriate courses.

Recognition of EnglishScore results

Today, over 1,000 organisations around the world, representing a diverse set of industries, use and recognise EnglishScore certificates. Employers have used EnglishScore as part of the process of recruitment and screening of potential staff and for upskilling their existing workforces. Universities have used EnglishScore as part of their admissions and placement procedures, and also as an exit credential for graduates entering the workforce.

To learn more about EnglishScore, please visit www.englishscore.com.

Test delivery

EnglishScore is an on-demand test and is administered and proctored through a mobile device. Test-takers download an app (available on iOS and Android), register their details and then take the test on their phone. It is free to access, and results are typically delivered within 24 hours of completing the test, with the option to purchase a certificate on completion of the test. More details on proctoring and other security features are detailed in the EnglishScore Security Report.

Speaking Test

The EnglishScore speaking assessment complements and supports the Core Skills Test and is designed to measure a test-taker’s speaking proficiency in everyday and workplace scenarios. It is delivered through the EnglishScore app and requires the test-taker to complete the EnglishScore Core Skills Test first.

II. Test design

II. Test design

a. Test development

EnglishScore was developed in association with the Centre for Research in English Language Learning and Assessment (CRELLA) at the University of Bedfordshire, UK (www.beds.ac.uk/crella). CRELLA is widely recognised as the UK's leading centre for language assessment research.

Test structure

Like the Core Skills Test, the EnglishScore Speaking Test is informed by the sociocognitive model of language use originating in Cyril Weir's *Language Testing and Validation* (2005). In this model, both context and cognitive validity contribute to spoken performance, and these, along with other factors such as test-taker characteristics, are considered when designing test tasks.

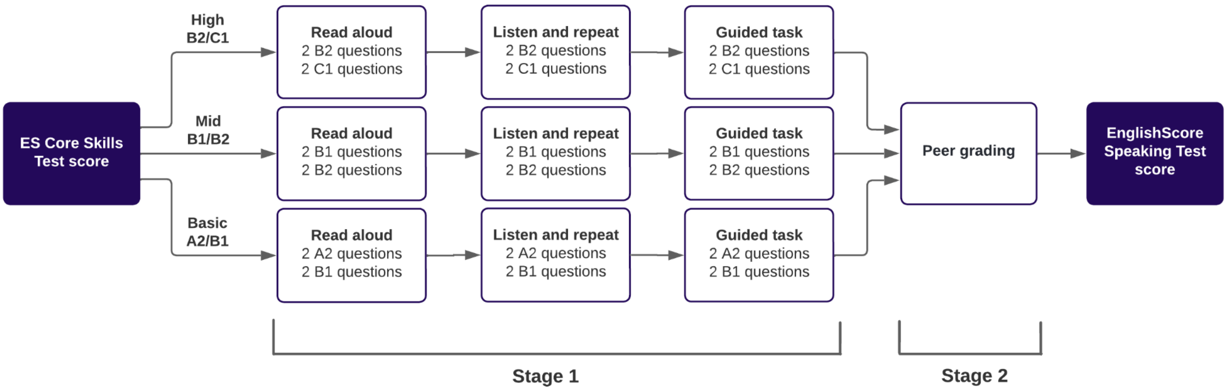
The EnglishScore Speaking Test is a single test, assessing a test-taker's spoken proficiency. The difficulty of the test items changes according to three branches or levels:

Basic (Breakthrough +):	CEFR levels A2 and B1
Mid (Threshold +):	CEFR levels B1 and B2
High (Vantage +):	CEFR levels B2 and C1

A test-taker is assigned one of the three levels based on their performance in the EnglishScore Core Skills Test which assesses grammar, vocabulary, listening and reading ability. This approach provides both an efficient and positive testing experience for test-takers, ensuring that they are not presented with items that are too difficult or too easy for them.

The speaking test consists of two main stages. Stage 1 includes these three sections: **Read aloud, Listen and repeat, and Guided task.** Stage 2, **Peer grading**, requires test-takers to listen to and rate other test-takers' responses.

__Figure 1.__ The EnglishScore Speaking Test stages.


ES Core Skills Test score
High B2/C1
Mid B1/B2
Basic A2/B1
Read aloud 2 B2 questions 2 C1 questions
Listen and repeat 2 B2 questions 2 C1 questions
Guided task 2 B2 questions 2 C1 questions
Read aloud 2 B1 questions 2 B2 questions
Listen and repeat 2 B1 questions 2 B2 questions
Guided task 2 B1 questions 2 B2 questions
Read aloud 2 A2 questions 2 B1 questions
Listen and repeat 2 A2 questions 2 B1 questions
Guided task 2 A2 questions 2 B1 questions
Peer grading
EnglishScore Speaking Test score
Stage 1
Stage 2

Additionally, the combined scores derived from AI scoring and peer grading are reported in the EnglishScore app and certificate (see Section ‘Scoring model’).

More information can be found in the ‘Assessment stages and tasks’ section.

b. Domains assessed

The speaking assessment tests spoken English in the personal, public and occupational domains, with an emphasis on workplace English at the higher CEFR levels. The test does not require specialist knowledge of particular domains, and questions are based on commonly accessible, everyday and work-based topics such as conversations with colleagues, the use of public transport and interactions with friends and family.

As the CEFR level targeted by the item increases, the domain of the input changes: at A2 level, the input mainly relates to the *personal* domain; at B1 level, the *public* domain; and at B2 and C1 levels, the *occupational* domain. At lower levels, the input is mostly very *concrete* and *familiar*, e.g. *about where people live, people they know and things they have or common objects around them*, progressing to less familiar and more abstract topics at the higher levels, e.g. *a wide range of both familiar and unfamiliar topics normally encountered in personal, social or vocational life.*

Test delivery

EnglishScore is an on-demand test and is administered and proctored through a mobile device. Test-takers download an app (available on iOS and Android), register their details and then take the test on their phone. It is free to access, and results are delivered within 24 hours from the end of the test, with the option to purchase a certificate. More details on proctoring and other security features are detailed in the EnglishScore Security Report.

III. The content of EnglishScore

III. The content of EnglishScore

a. Overview

Stage 1 – The speaking assessment

This section covers Stage 1, the speaking assessment stage. Stage 2 (the peer grading stage) is described in Section D.

Each of the three levels (A2/B1, B1/B2, B2/C1) in Stage 1 contains the following item types. The number of items in each level is the same; however, the items differ in their difficulty levels:

Read aloud: four items
Listen and repeat: four items
Guided task: four items

Within each branch, six items are targeted at lower and six at higher CEFR levels.

Connection to language use
The tasks are similar to speaking activities that are commonly encountered in personal, public and general occupational settings around the world.

Instructions
The instructions are given in English.

Timing
In the test, it takes around 15–20 minutes to complete Stage 1 and 10–15 minutes to complete Stage 2.

What is the input?

At the start of the test, the test-taker is asked to confirm that the microphone, speakers and camera are working as expected.

The input

Table 1 describes the input features. As the CEFR level targeted by the item increases, the domain of the input changes: at A2 level, the input mainly relates to the *Personal* domain; at B1 level, the *Public* domain; and at B2 and C1 levels, the *Occupational* domain.

Communication themes
Communication themes may include *personal identification/house, home, environment/daily life/free time/entertainment/relations with other people/shopping/food and drink/places and weather*.

Sources of input

Item content is prepared by item writers specially for EnglishScore. Items are written to reflect the spoken features test-takers would encounter in the relevant domains. Writers use English Profile (www.englishprofile.org) Reference Level Descriptions[3] for English to guide the language difficulty of the items.

Nature of input

At lower levels, the input is mostly very *concrete* and *familiar*, e.g. *about where people live, people they know and things they have, common objects around them*, progressing to less familiar and more abstract topics at the higher levels, e.g. *a wide range of both familiar and unfamiliar topics normally encountered in personal, social or vocational life.* For spoken input, the delivery is clearly articulated at a natural rate.The input recordings involve only one speaker.

Difficulty level of the input

The input is likely to be comprehensible to a language learner at a CEFR level just below the target level. In other words, an item targeting A2 is intended to be comprehensible at A1, an item targeting C1 is intended to be comprehensible at B2+, etc.

Test items for each item type are presented in approximate order of difficulty.

[3] See Cambridge University Press (2015) for further information. English Profile helps teachers and educationalists to understand what the CEFR means for English. It describes what aspects of English are typically learnt at each CEFR level.

Table 1. *The assessment input.*

	CEFR level	Example focus	Input length and level	Domain	Nature of info, topic familiarity
Read aloud	**A2**	Read out phrase on screen	Around 10 words	Personal and public	Concrete, familiar
	B1	Read out phrase on screen	Around 14 words	Personal and public	Concrete, familiar
	B2	Read out phrase on screen	Around 18 words	Public and professional	Mix of concrete and abstract, mix of familiar and unfamiliar
	C1	Read out phrase on screen	Around 22 words	Public and professional	Abstract, unfamiliar
Listen and repeat	**A2**	Listen to speaker, repeat heard phrase	Around 6 words	Personal and public	Concrete, familiar
	B1	Listen to speaker, repeat heard phrase	Around 9 words	Personal and public	Concrete, familiar
	B2	Listen to speaker, repeat heard phrase	Around 12 words	Public and professional	Mix of concrete and abstract, mix of familiar and unfamiliar
	C1	Listen to speaker, repeat heard phrase	Around 14 words	Public and professional	Abstract, unfamiliar
Guided task	**A2**	Tell a friend about the weather	-	Personal and public	Concrete, familiar
	B1	Give directions	-	Personal and public	Concrete, familiar
	B2	How can we help the environment at work?	-	Public and professional	Mix of concrete and abstract, mix of familiar and unfamiliar
	C1	Talk about a problem at work	-	Public and professional	Abstract, unfamiliar

Read aloud

In this item type, test-takers see a short sentence on screen and read it out loud. Sentence word count ranges from around 10 to 22 words. The phrase required to be repeated will increase in length and complexity dependent on the CEFR level of that item.

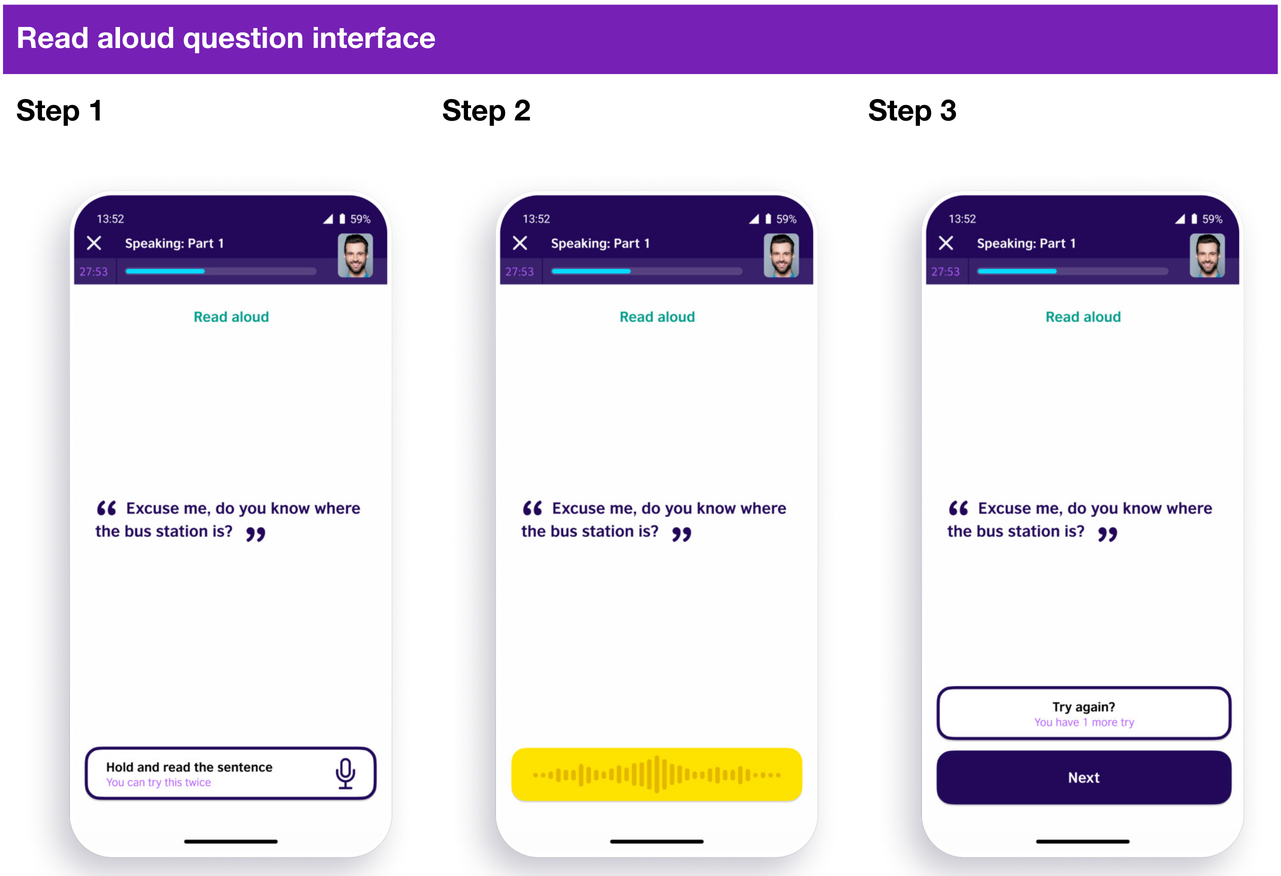
All instructions for the Read aloud are given in writing which remains on screen during the assessment. Test-takers can re-record a response if they wish (but cannot listen to their first attempt).

Test-takers must read, understand and produce the sentence, allowing assessment of a test-taker's pronunciation and fluency in spoken English.

Test steps
Figure 2 shows the stages of the Read aloud question flow for the test-taker.

Step 1: First, the test-taker is instructed to read the phrase on the screen and then record their responses.
Step 2: The test-taker records their responses, reading the phrase aloud.
Step 3: The test-taker has the option to record their responses again or move on to the next question. Note: They are not allowed to listen to their first attempt before deciding whether to try recording again or not.

Figure 2. *Read aloud.*



Read aloud question interface

Step 1

13:52 59%

Speaking: Part 1

27:53

Read aloud

“Excuse me, do you know where the bus station is?”

Step 2

13:52 59%

Speaking: Part 1

27:53

Read aloud

“Excuse me, do you know where the bus station is?”

Step 3

13:52 59%

Speaking: Part 1

27:53

Read aloud

“Excuse me, do you know where the bus station is?”

Listen and repeat

In this item type, test-takers hear a short sentence and repeat it out loud verbatim. Item word counts range from 5 to 16 words. The phrase required to be repeated will increase in length and complexity dependent on the CEFR level of that item. Test-takers can listen to the sentence only twice.

Test-takers must listen to, understand and produce the sentence, allowing assessment of a test-taker's pronunciation and fluency in spoken English.

Listen and repeat items require test-takers to organise the speech into linguistic units. For test-takers, mastery of sentence structure is important, as high-proficiency test-takers can repeat a long string of words because they are, in general, more familiar with English sentence structure. Additionally, the ability to repeat full sentences can show test-takers' fluency and pronunciation abilities in spoken English.

Test instruction prompts are delivered in a conversational manner, using a range of different accents. Spoken features to provide meaning are also used in the items. These features may include question forms, conditionals, uncertainty and exclamations.

Test steps

Figure 3 shows the stages of the Listen and repeat question flow for the test-taker.

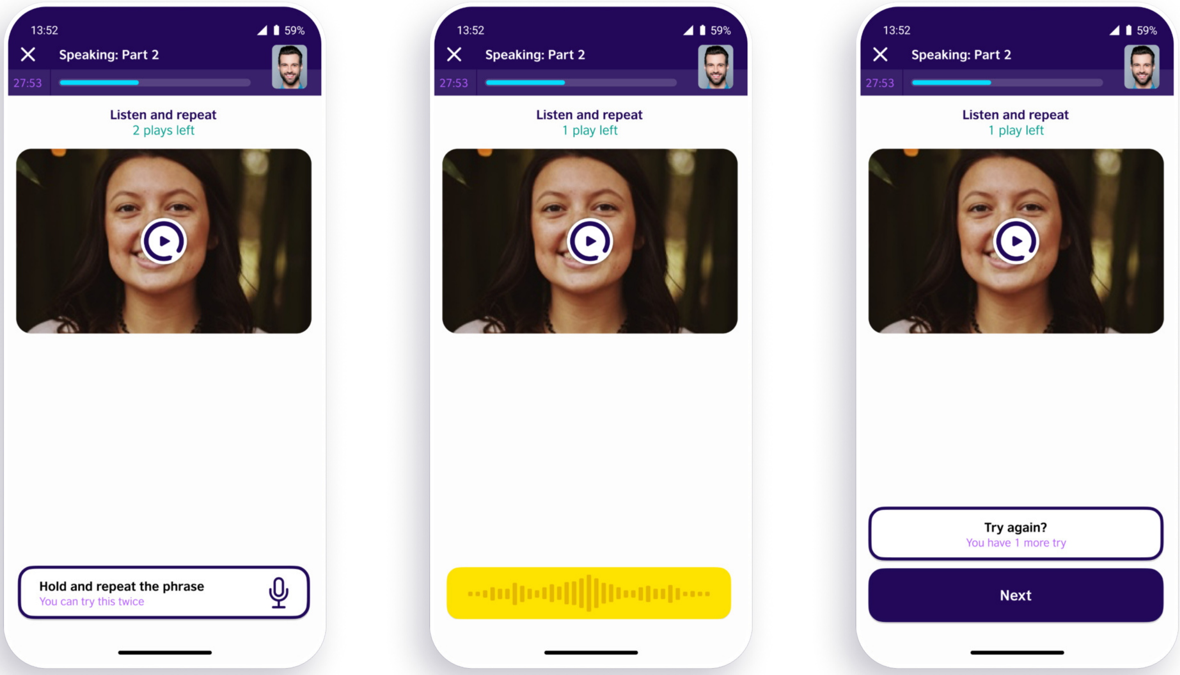
Step 1: The test-taker is shown a video of a person speaking a phrase; they are able to watch this video twice. The test-taker is then prompted to record their responses.
Step 2: The test-taker records themselves (audio only) repeating the phrase.
Step 3: The test-taker has the option to record their responses again or move on to the next question. Note: They are not allowed to listen to their first attempt before deciding whether to try recording again or not.

__Figure 3.__ Listen and repeat.

Listen and repeat question interface

Example video phrase: ‘I'll let you know when I get the exact date.’

Step 1 **Step 2** **Step 3**



Guided task

In this item type, test-takers hear a short question asking for information or an opinion and are asked to respond. There are **two versions** of this item type: textual and visual. **A textual guided task** presents a question to the test-taker and asks them to respond to it for a given number of seconds depending on level. **A visual guided task** is similar, but there is also graphical input such as an image or a table to guide the response. As they are giving their response, written prompts on screen remind the test-taker of the topic and questions.

This task type assesses extended speaking turns, measuring overall comprehensibility and communication (task achievement), such as providing relevant and supporting detail in responses.

Test prompts are delivered in a conversational manner, using a range of different accents.

Test steps

Figures 4 and 5 show the stages of the Guided task question flow for the test-taker.

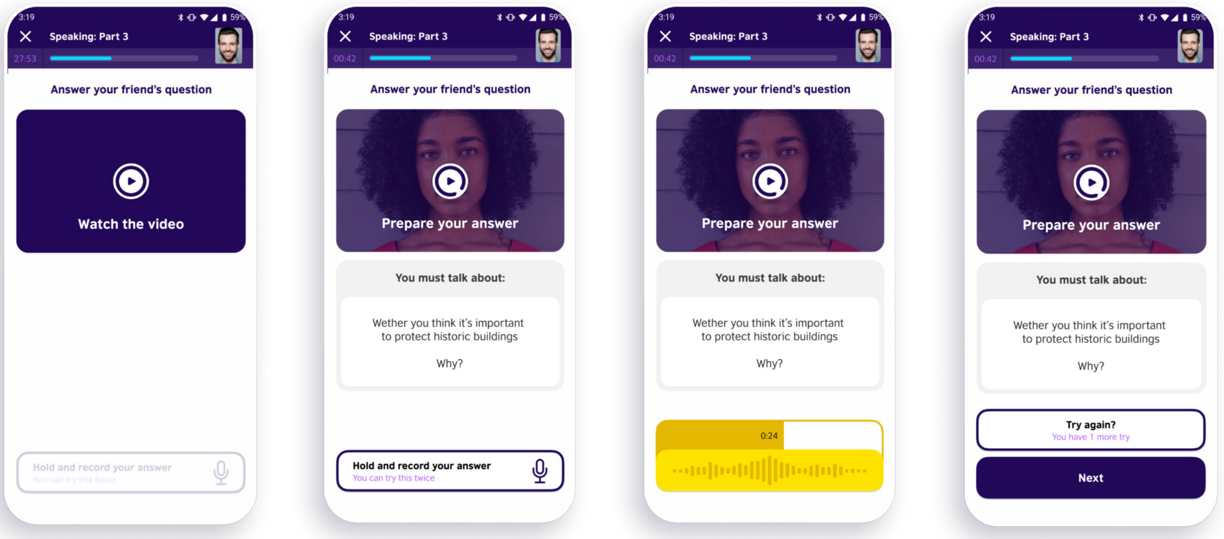
Step 1: The test-taker is shown a video of a person presenting a topic. They can watch this video twice.
Step 2: Where appropriate, the test-taker is shown reference material to use in completing the task and prompted to record their reply.
Step 3: The test-taker records their responses. A maximum talking time is enforced for this question type: A2 and B1 = 30 seconds, B2 and C1 = 45 seconds.
Step 4: The test-taker has the option to record their responses again or move on to the next question. Note: They are not allowed to listen to their first attempt before deciding whether to try recording again or not.

Figure 4. Guided task – textual question.

Guided task textual question interface

Example video script: 'We have lots of historic buildings in my town. Do you think it's important to protect historic buildings?'

Step 1 **Step 2** **Step 3** **Step 4**



You must talk about:

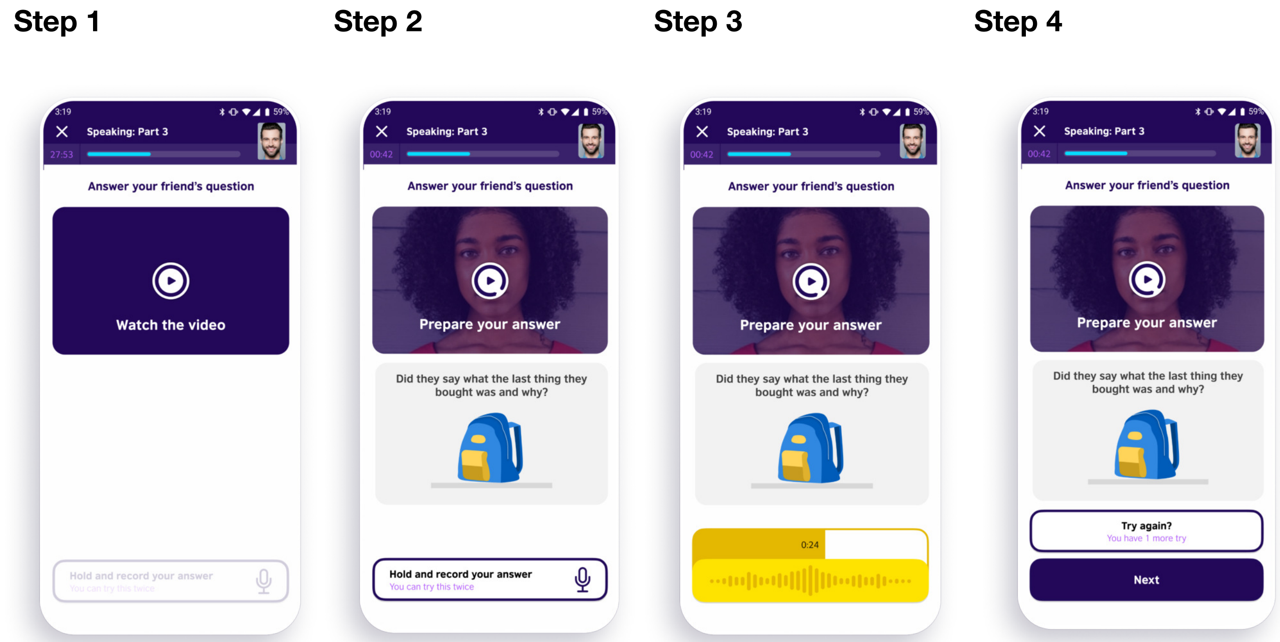
Wether you think it's important to protect historic buildings

Why?

__Figure 5.__ Guided task – visual question.

Guided task visual question interface

Example video script: ‘I went shopping yesterday and bought a new bag. What was the last thing you bought?’



Step 1

Speaking: Part 3

Answer your friend's question

Step 2

Speaking: Part 3

Answer your friend's question

Did they say what the last thing they bought was and why?

Step 3

Speaking: Part 3

Answer your friend's question

Did they say what the last thing they bought was and why?

Step 4

Speaking: Part 3

Answer your friend's question

Did they say what the last thing they bought was and why?

b. Writing the assessment material for EnglishScore

Test writer qualifications

EnglishScore writers are teachers of English with a teaching qualification such as a Masters’ Degree or Diploma in English Language Teaching and a minimum of five years’ experience as teachers of English. They are also familiar with the CEFR and able to write items to the different CEFR levels. Before being accepted for training, writers complete a qualifying item-writing task.

Test writer training

All writers are given an induction programme to the test, where they are introduced to the test specifications and practise writing assessment material. Writers regularly participate in review meetings and are required to complete a training course every three years to continue working as contributors to the EnglishScore assessment.

Test writer guidance

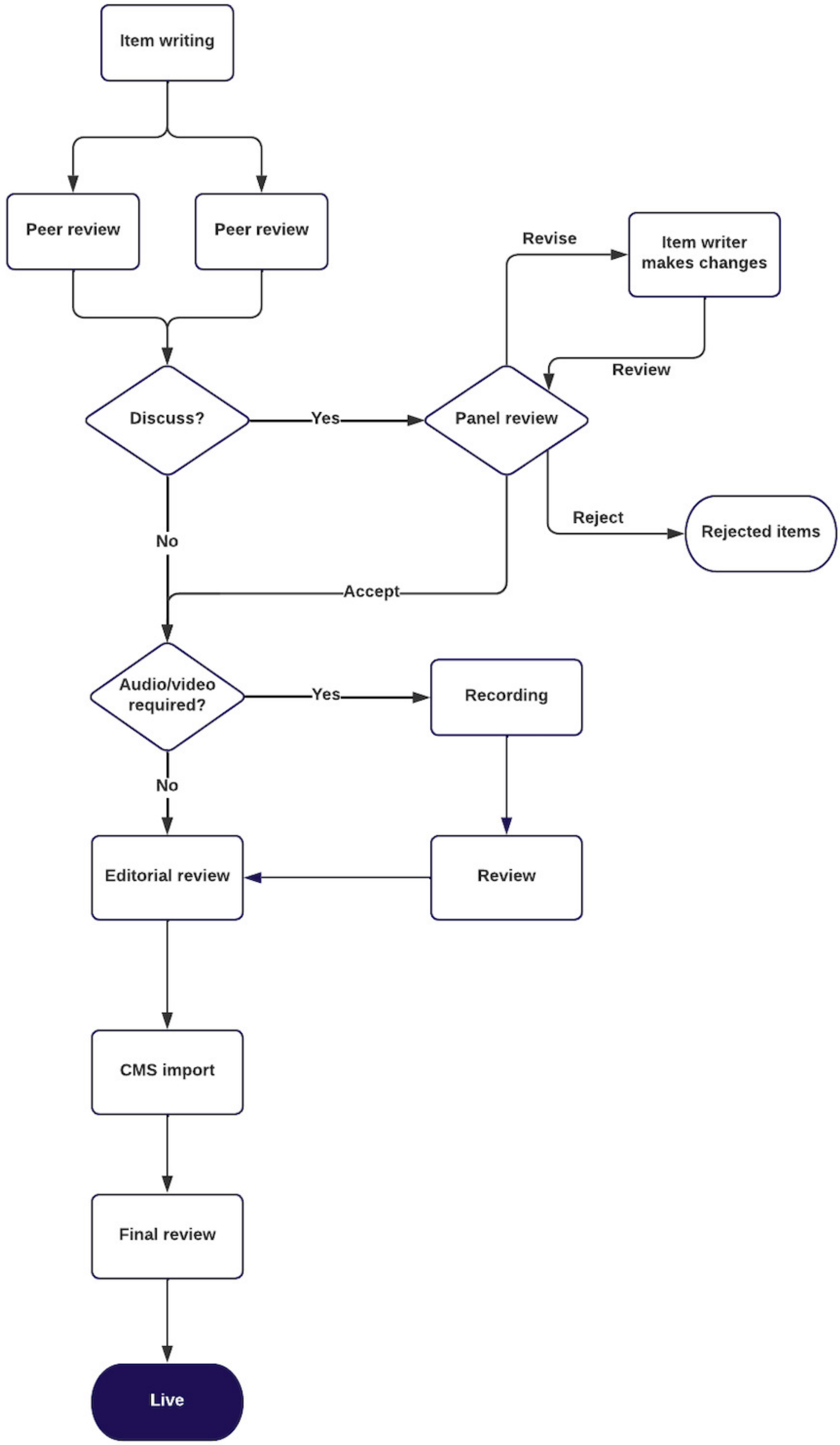
To ensure that the content follows the developers' intentions and that it is parallel across different versions of EnglishScore, when preparing material, the writers follow detailed writer guidelines for each section of the test. These include examples of good (and poor) tasks, wordlists, lists of grammatical structures and guidance on features such as text and recording length, rates of speech and complexity. There are self-assessment checklists that writers use to confirm that their work conforms to the guidelines before they submit it. Item writers are also instructed to use automated text tools (e.g. Vocab Profile, Text Inspector) to ensure vocabulary and grammatical structures of speaking prompts are at the appropriate level for test-takers. Additional training and feedback are provided to item writers by the EnglishScore Senior Assessment Manager as needed, and reviews are included in the item development process.

Test material development process

Test items are developed by a team of trained item writers in a series of item commissions throughout the year. To ensure the quality of items and the test as a whole, there is a standardised set of procedures that all items go through. This helps to ensure that test quality is maintained and that the test results are reliable and fair. To ensure consistency, the process mirrors that used for the EnglishScore Core Skills Test item development.

An overview of the item development process is provided below in Figure 6.

Figure 6. *EnglishScore Speaking Test item development process.*



Notes

- At any stage of the item development process, material may be accepted for the next stage, edited, returned to the writer for revision or rejected.
- Items are reviewed by item writers individually and as a panel and evaluated against the test specifications, assessing aspects such as item content, CEFR level and word count.
- All review decisions and feedback during the process are securely recorded for reference at a later date if needed.
- Final checks include an editorial review for proofreading and style checks, before being ingested into the secure item bank and individually reviewed and activated by the EnglishScore Senior Assessment Manager.

Once the items are live, weekly checks are conducted to ensure the items are performing as expected. Any items that fall outside the quality parameters are flagged, deactivated and removed from the test.

Item bank and test security

All the test items are stored in a secure item bank which includes the item content, item media and metadata (level, skill, etc.). Items from this database are selected to create a large number of unique test forms. The item bank is large enough to ensure there is minimal repetition of items across multiple test attempts by the same test-taker, which helps to maintain the security of the item bank.

The item bank and associated CMS are maintained by EnglishScore. Access to the item bank is restricted and controlled through a username and password. All changes to item content are logged with date/time/username, with access permissions regularly reviewed.

Additional details on test security are available in the *EnglishScore Security Report*.

Taking account of test-taker needs

EnglishScore takes account of the diversity of the test-taking population by collecting data about their location and by asking test-takers about their motivation for learning English.

The test material is designed for young adult and adult learners of English (aged 16 and over) and aims to avoid any bias associated with gender, nationality or ethnic identity. These issues are addressed in the guidelines for item writers and considered as part of the review process. In addition, items are checked to ensure that they do not include controversial topics, do not require specialist knowledge and that they are culturally neutral, i.e. do not require knowledge of a particular culture or country to be answered correctly. This ensures test fairness for all test-takers around the world.The test interface is designed to be accessible to colourblind users.

IV Scoring

a. Scoring

The EnglishScore Speaking Test is designed to be an accurate, reliable measure of a test-taker's speaking ability in the global workplace. To achieve this, we use a blend of automated scoring and peer grading to calculate a speaking score that reflects how well a test-taker can communicate with people from a variety of different backgrounds, cultures and English language levels.

Stage 1: Assessment tasks

Read aloud and Listen and repeat items are scored using a technical solution, and the Guided task items are scored using peer gradings.

Read aloud: marking

This item type is evaluated by automated AI scoring for:

- Pronunciation – phoneme and stress accuracy
- Fluency – rate of speech and pausing.

Listen and repeat: marking

This item type is evaluated by automated AI scoring for:

- Pronunciation – phoneme and stress accuracy
- Fluency – rate of speech and pausing.

Guided task: marking

This item type is evaluated by peer markers for:

- Communication
- Overall comprehensibility.

Stage 2: Peer grading

Stage 2 of the assessment asks test-takers to rate other test-takers' Guided task responses, which is compulsory. Test-takers are prompted to rate other test-takers' responses honestly, and they are aware that their answers are also rated by other peers. A test-taker is marked by approximately 16 unique peers, all at similar proficiency levels. Sourcing a wide range of peer opinions reduces bias and ensures a more accurate and reliable measure of a test-taker's proficiency. In addition, EnglishScore regularly reviews peer rater reliability through controlled sittings to ensure that peer raters are scoring consistently.

Test-takers will peer-mark other test-takers from the same test level, e.g. a test-taker who completed the ‘mid-level’ item will mark other test-takers who completed the same level, but not the low or high level. For instance, test-takers peer-mark the Guided task question types only. The Read aloud and Listen and repeat item types are scored using automated AI scoring.

Grading steps

Figures 7 and 8 show the stages of peer grading for the test-taker.

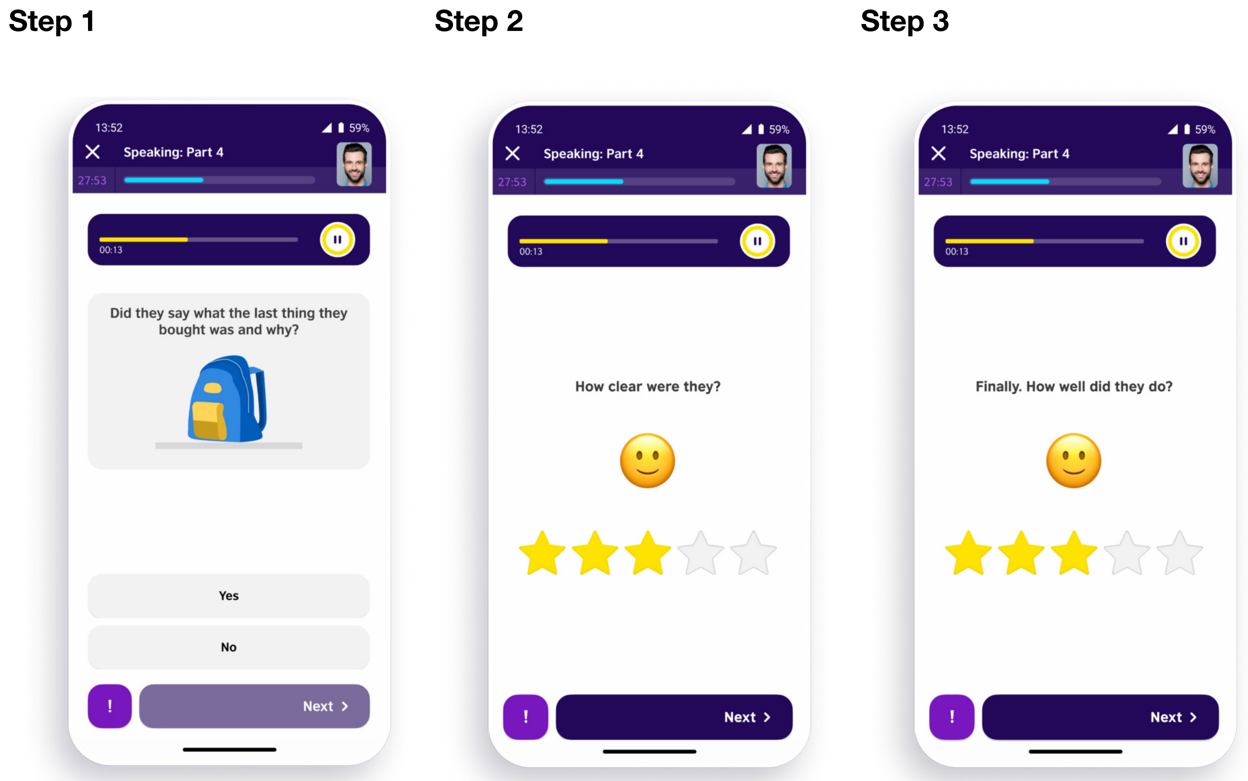
Step 1: The rater can listen to a test-taker’s response to a Guided task. They can listen to this as many times as they wish. The first grading question relates to task completion – which is a binary (yes/no) or trinary question (yes, one/yes, both/no, none), depending on the item.
Step 2: The second grading question asks the rater to choose from a 5-star scale for clarity of the ideas.
Step 3: The third question asks the rater how well the test-taker did on a 5-star scale.

On each grading screen, the rater can listen to the response as many times as they need to.

Figure 7. *Guided task*

Guided task visual grading interface

Example video script: ‘I went shopping yesterday and bought a new bag. What was the last thing you bought?’


Step 1

Did they say what the last thing they bought was and why?

Step 2

How clear were they?

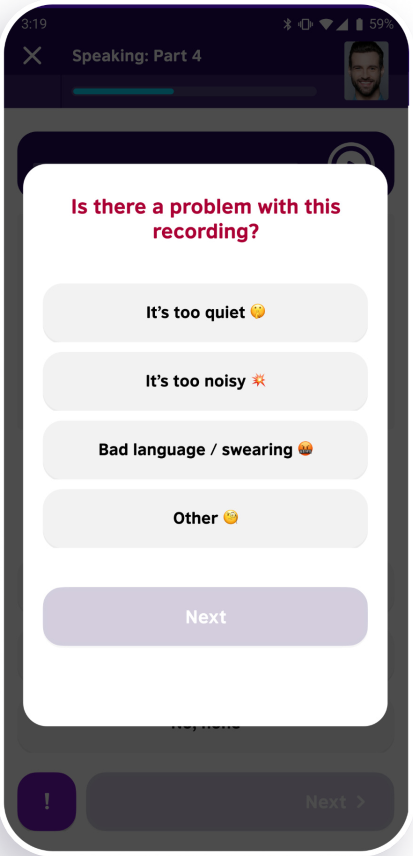
Step 3

Finally. How well did they do?

Flagged items while peer scoring

During the peer grading, test-takers have the option to flag items that they feel they can’t score. This may be because the audio quality is too poor, the content is inappropriate or some other reason (Figure 9). If a response is flagged, the item is skipped and the rater is presented with the next item.

__Figure 9.__ Example for flagged options.



Is there a problem with this recording?

b. Score reporting

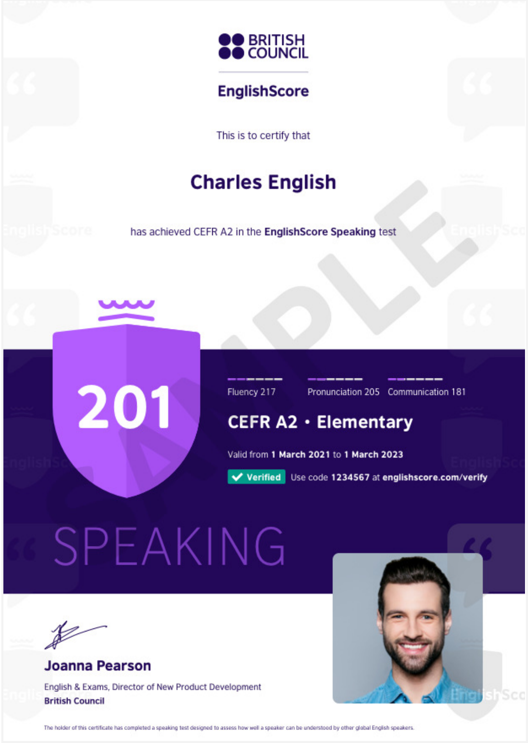
On completing the EnglishScore Speaking Test, the test-taker is provided with an onscreen report stating their overall ‘EnglishScore’ speaking score, with separate score breakdowns for pronunciation, fluency and communication.

Estimated correspondences to CEFR level and estimates of correspondences to IELTS scores and the Cambridge English Scale are also provided.

EnglishScore range	CEFR level
0–199	Below A2
200–299	A2
300–399	B1
400–499	B2
500–599	C1

Test-takers have the option of purchasing a certificate as a record of their score. Each certificate includes the test-taker’s name, a photograph of the test-taker taken during the administration, a verification ID for use by employers or other score users and scores for overall speaking and the subskills (pronunciation, fluency and communication).

This is an example Speaking Test certificate:



BRITISH COUNCIL

EnglishScore

This is to certify that

Charles English

has achieved CEFR A2 in the **EnglishScore Speaking** test

SAMPLE

201

Fluency 217 Pronunciation 205 Communication 181

CEFR A2 • Elementary

Valid from **1 March 2021** to **1 March 2023**

✓ Verified Use code **1234567** at **englishscore.com/verify**

SPEAKING

Joanna Pearson

English & Exams, Director of New Product Development

British Council

The holder of this certificate has completed a speaking test designed to assess how well a speaker can be understood by other global English speakers.

Pass marks

There are no pass marks for EnglishScore. Scores are reported in relation to the Common European Framework of Reference for Languages (CEFR) from A2 to C1. Estimates of a test-taker's CEFR level are based on their success in responding to material targeting each level. Further work will be undertaken to set standards in relation to the CEFR and to performance in other tests.

EnglishScore scale

The EnglishScore is a numeric, granular scale which measures English language proficiency from 0 to 599. It builds on the Common European Framework of Reference (CEFR) by showing finer gradations within a learner's CEFR level and can therefore help to measure gradual improvements in a test-taker's English level across the different skills. As well as providing useful and motivating feedback to test-takers, it also gives teachers and other decision makers a more detailed understanding of test-takers' strengths and weaknesses.

Time for results

Speaking score results are typically reported within 24 hours of a test-taker completing their peer grading.

Reporting

At the end of the test, the test-taker's speaking ability is reported as a speaking score from 0 to 599 on the EnglishScore scale, as well as the corresponding CEFR level. In addition, a breakdown of speaking subskills (pronunciation, fluency and communication) is also provided, as well as a set of can-do statements to provide additional context to the reported test score.

c. Scoring model

The EnglishScore Speaking Test scoring model consists of two components:

- Automated scoring using AI technology – this recognises speech patterns of test-takers and scores different traits such as phoneme accuracy, stress, rate of speech and pausing ratios

- Peer grading – test-takers score each other on different attributes, including task completion and overall comprehensibility.

Both data sources feed into the scoring model to give an overall speaking score, plus subskill scores for pronunciation, fluency and communication.

Scoring model design

A key principle for the scoring model was to ensure alignment with expert raters, i.e. the EnglishScore speaking model is designed to score as an expert human rater would. As part of the scoring model development, over 8,000 spoken responses were collected from a range of test-takers at different CEFR levels and from different countries around the world. These responses were then rated by expert markers to provide scores used to build the scoring model. The expert raters' evaluations of performance data also allowed us to build and refine the analytic rating scale descriptors. The rating scale criteria are *pronunciation, task fluency* and communication (*task achievement)*. The score levels range between 0 and 6, from 'no evidence' to 'proficient' (see rating scale in Appendix 1).

To provide the expert scores, a group of experienced speaking raters were recruited, trained and certified to use the EnglishScore speaking descriptors. Each spoken response was then rated independently by at least two experts for pronunciation, fluency and communication (task achievement). Each expert rater graded 12 responses per test. Overall, 6 raters rated 650 test-takers' spoken performances. An average of the two ratings was then used to build the scoring model. Where the two rater scores were significantly different, a third rating from a senior examiner was used to determine the final score. As part of the rating activity, the raters were given calibration tasks, and spot checks were carried out by the English Score Senior Assessment Manager.

The robustness of the scoring model was evaluated by comparing the correlation coefficient with the expert raters. The model went through several iterations, combining and weighting a range of automated AI scoring and peer gradings to arrive at a model that correlates strongly with experts. The current model has a strong correlation of 0.85.

Speaking subskills

As well as an overall score, subskills are also reported in the app and on the certificate. These provide a more detailed breakdown of a test-taker's strengths and weaknesses.

Pronunciation – can the test-taker produce speech that is easily understandable to most speakers of the language? There is no particular desired accent; the only criterion is that it should be globally comprehensible. Factors such as appropriate phoneme pronunciation and use of appropriate stress lead to a higher score in this subskill.

Fluency – can the test-taker produce speech that is smooth and at a natural speed? Factors such as a constant rate of speech with appropriate pausing will lead to a good score in this domain.

Communication (task achievement) – can the test-taker produce an answer that is relevant to the prompt and contains additional detail and supplementary information where appropriate?

d. Evidence about the score reliability and accuracy

To ensure that the Speaking test scores are valid and reliable, EnglishScore regularly reviews test performance using a range of methods and metrics. These provide valuable insights into the reliability and validity of the scoring methods and help us identify any necessary improvements or adjustments to the test.

Test-retest reliability

This measures the correlation between scores when a test-taker attempts the test more than once. A valid and reliable test should produce scores which are consistent and only vary slightly from one test attempt to the next (assuming no learning has taken place between attempts). The test–retest data in the EnglishScore Speaking Test is calculated on a daily basis. The test–retest reliability coefficient for the EnglishScore Speaking Test was 0.81 for the tests taken between 1 February 2023 and 1 April 2023, where the same test-taker had taken the test twice in a month. The total number of test repetitions used to calculate the coefficient was approximately 57,000 (*N* = 57,134).

AI-scored items

As part of the ongoing test quality analysis, a sample of test data was analysed in March 2023 to monitor the effectiveness of the items as well as test-taker performances. A total of 3,003 EnglishScore test-takers responded to Speaking items and were scored on 2 task types and 3 features or traits.

The task types were:

Listen and repeat

Read aloud

These task types were scored using AI/tech scoring approaches. These are the traits measured in each task type:

Pronunciation – phoneme and stress accuracy

Fluency – rate of speech and pausing.

For the purposes of this analysis, each feature was scored on a scale of 0 to 6. The test at easy (A2/B1), medium (B1/B2) and high (B2/C1) levels was analysed using the WINSTEPS Rasch software program. As WINSTEPS does not recognise decimals, all scores are multiplied by 10 so that test-taker scores range from 0 to 60. In other words, all scores were rounded to two decimal places prior to the analysis.

Overall, the test reliably separated item and test-taker levels, and the fit indexes were as expected by the model.

In terms of fit, the infit and outfit MnSq values were reviewed to see if the scores meet the Rasch model expectations. These measures show the degree to which a particular test question contributes to the overall test score. By examining the infit and outfit MnSq values, we can identify problematic questions or items and make necessary updates or adjustments to improve the overall quality and validity of the test.

By setting acceptable levels (Linacre, 2012), around 90% of the sample data fitted the model expectations. The indexes showing underfit or overfit and unsuitable difficulty or item correlation values were analysed, and necessary updates were conducted to improve the test quality.

To ensure the test quality, test performance is continually monitored and reviewed by the EnglishScore Senior Assessment Manager. Figure 10 below provides a detailed item review workflow, showing how those flagged items are reviewed to improve the accuracy of the scores as well as the quality of the items.

Peer-scored items

Rationale

The scoring model relies on inputs from other test-takers or ‘peers’, which are then combined with automated AI scoring. The purpose of using peers in the scoring is twofold: automated scoring cannot currently assess task completion and overall spoken production to a reliable and accurate degree, and peer grading reflects what test-takers are expected to do in a real-world workplace setting.

As part of the test design and scoring model, there are several features to ensure the accuracy and reliability of the peer grading:

- Peer raters will only assess other test-takers in the same level – this means they are approximately the same English level, which helps to reduce bias when scoring.

- Peer grading questions are straightforward – the language used in the peer grading questions is at A2 or lower, meaning they are understood by test-takers (e.g. ‘*how clear were they?*’). Required responses are also straightforward and easily understood.

- A minimum of 16 peer raters are used to derive a test-taker’s score. In contrast to many human-rated speaking tests that may only rely on 1 or 2 raters, the data obtained from 16 peer raters’ scoring is believed to be more robust.

- During the peer grading, test-takers have the option to flag items that they feel they can’t score. This may be because the content is inappropriate or some other reason. If a response is flagged, the item is skipped and the rater is presented with the next item.

- The other measure includes identifying and removing the ‘rogue’ peer raters. Rogue graders refer to the peer raters who constantly give the same score regardless of the performance. These rogue peer raters are detected and their scores are removed prior to the score calculation. This ensures that those rogue raters’ marks do not affect any test-taker’s scores.

- The scoring algorithm also allows us to detect the outlier peer raters whose scores significantly differ from the group. Implementing this step prior to the score calculation ensures that the outlier peer raters’ markings do not affect the score accuracy.

To ensure reliability and consistency, the test performances and peer scoring results are continually monitored. The data is regularly reviewed by the EnglishScore Senior Assessment Manager.

This report will be regularly updated as we continue monitoring the quality of our test and collecting further validity evidence.

Item review and health

The item review process helps to ensure that the items are valid, reliable and appropriate for the intended population. By identifying and removing poorly constructed or ambiguous items, the review process helps to reduce measurement error and increase the precision of language measurements.

In the item review process, the items flagged according to the Rasch parameters are identified and reviewed by our team of language assessment experts and item-writing specialists. Reviewing the items as well as test-taker performances (speaking scripts) helps us to improve the items.

There are three main stages in the review process:

Pre-review stage:

At this stage, the flagged items are identified based on defined Rasch parameters such as item difficulty and item fit. The flagged items are then listed and prepared for review. The preparation may include gathering additional information such as sample audio responses.

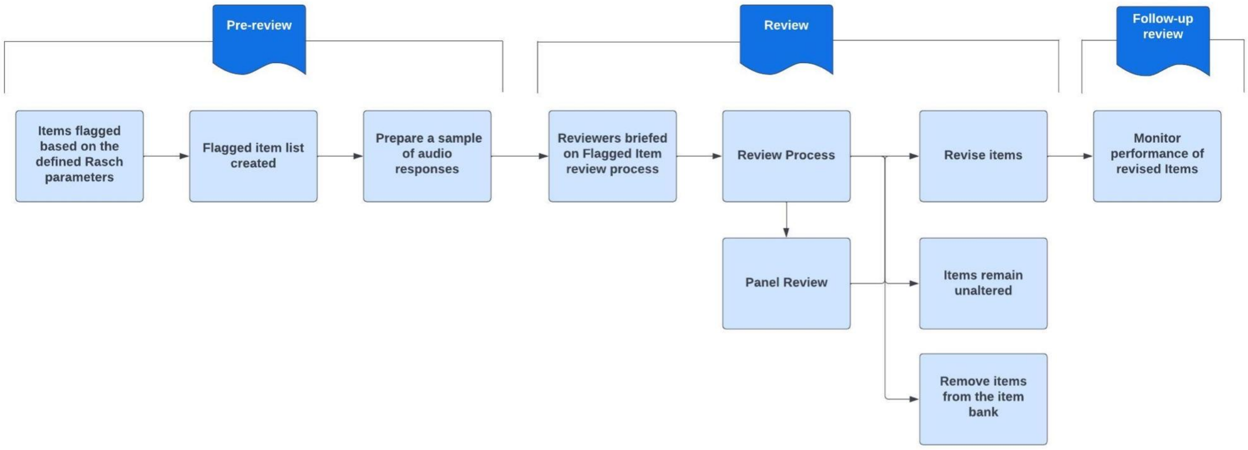
Review stage:

At this stage, the flagged items are reviewed by experts who evaluate the items for quality and appropriateness. The review process may involve revising the items, leaving them as they are or removing them from the item bank. Additionally, a panel review is conducted where multiple experts review the flagged items and provide feedback on their quality and appropriateness. This helps to ensure that the items are of high quality and meet the necessary criteria for use in the language assessment.

Follow-up review stage:

After the initial review, the updated items are monitored to ensure that they meet the necessary criteria and are functioning as intended. This stage may involve further review or modifications to the items as needed.

***Figure 10**. Analysing flagged responses in speaking test: item review workflow.*


Pre-review
Review
Follow-up review
Items flagged based on the defined Rasch parameters
Flagged item list created
Prepare a sample of audio responses
Reviewers briefed on Flagged Item review process
Review Process
Panel Review
Revise items
Items remain unaltered
Remove items from the item bank
Monitor performance of revised Items

The implementation of a rigorous item review process results in more precise evaluations of language proficiency and better-informed judgements about the language abilities of individuals taking the EnglishScore test. In addition, the findings from the reviews inform the item development process, enabling the refinement of future test items and enhancing the overall quality of the EnglishScore Speaking Test.

Appendix 1. Rating scale for EnglishScore Speaking.

EnglishScore speaking descriptors v1.0					
Pronunciation (intelligibility, clarity and prosodic features)		**Fluency** (speed, hesitations and pauses)		**Communication (task achievement)** (task attempts and relevance)	
6 **PROFICIENT**	● Speech is immediately and clearly understood. ● All vowels and consonants are generally produced with clarity and precision with appropriate assimilations. ● Prosodic features such as word and sentence stress, intonation and rhythm are used appropriately and to convey finer shades of meaning.	**6** **PROFICIENT**	● Speech is smooth, with natural pacing. ● There are minimal hesitations, repetitions or false starts. ● Any pauses, repetitions or false starts are related to accessing ideas and not language.	**6** **PROFICIENT**	● Attempts all parts of the task in full. ● Responses are fully relevant and appropriately extended and/or developed.
5 **ADVANCED**	● Speech is generally clearly understood. Where there are errors, these do not affect intelligibility. ● Vowels and consonant sounds are produced clearly. ● Stress is placed correctly in all high-frequency words, and sentence level stress is sometimes used to convey meaning.	**5** **ADVANCED**	● Speech is generally smooth, with acceptable pacing. ● There may be a few hesitations, repetitions or false starts. ● Pauses are infrequent and unlikely to be language-related.	**5** **ADVANCED**	● Attempts all parts of the task. ● Responses are generally relevant and appropriately developed.

4 **GOOD**	● Some systematic errors in sounds might render a few words unclear and may affect intelligibility. ● Most vowels and consonants are produced correctly. ● There is some use of prosodic features, such as stress and intonation to convey meaning, but not consistently.	**4** **GOOD**	● Speech has acceptable speed but may be uneven in patches. ● There may be some hesitations, but most words are spoken in continuous stretches of speech. There are few repetitions or false starts. ● Speech has no long pauses and generally sounds connected.	**4** **GOOD**	● Attempts all parts of the task. ● Most of the responses are relevant and appropriately developed, though there may be some ambiguity.
3 **INTERMEDIATE**	● Most of the speech is intelligible, but the listener may at times require effort to understand the speaker. ● Some consonants and vowels are systematically mispronounced. ● Prosodic features are present, but not always appropriately. Stress may be placed incorrectly in some words, and/or intonation may be inappropriate, which can cause confusion for the listener.	**3** **INTERMEDIATE**	● Speech may have uneven pacing and/or staccato at times, which might be distracting for the listener. ● Extended utterances may have some smooth multiple-word runs, although hesitations, repetitions or false starts are also present. ● Extended utterances may have some long pauses.	**3** **INTERMEDIATE**	● Attempts most of the tasks, though some of the minor aspects of the task may not be attempted. ● Responses are mostly relevant and developed to some extent, though they may not always be relevant or can be ambiguous.
2 **BASIC**	● The listener may have difficulty understanding between a third and a half of speech. ● Many consonants and vowels are systematically mispronounced. ● Use of stress, intonation and rhythm may be inappropriate and cause strain for the listener.	**2** **BASIC**	● Speech is slow and has irregular pacing, which can cause strain for the listener. ● Speech is uneven with poor grouping, staccato speech and multiple hesitations, repetitions and/or false starts. ● Extended utterances may have noticeably long pauses.	**2** **BASIC**	● Attempts some tasks but in a limited way. ● Responses may not be relevant or developed beyond simple explanations.

1 **LIMITED**	● The listener may find most of the speech unintelligible. ● Most consonants and vowels are mispronounced or omitted, which causes severe strain for the listener. ● Use of stress, intonation and rhythm is largely inappropriate. Several words may have the wrong number of syllables. There is little to no control of intonation or rhythm.	**1** **LIMITED**	● Speech is very slow, with little grouping of words, which makes the message difficult to follow. ● There are multiple hesitations, pauses, false starts and/or reformulations. ● Most words are produced in isolation, and there may be multiple long pauses.	**1** **LIMITED**	● May attempt some parts of the task, but done very simply. ● Responses may be tangential or unrelated due to incomprehension of the task or a lack of language.
0 **NO EVIDENCE**	● No language or only a few isolated words produced. ● Response is completely off-topic, non-English or unintelligible. ● Poor audio quality means response cannot be scored.	**0** **NO EVIDENCE**	● No language or only a few isolated words produced. ● Response is completely off-topic, non-English or unintelligible. ● Speech is too short to reliably assess fluency. ● Poor audio quality means response cannot be scored.	**0** **NO EVIDENCE**	● No language or only a few isolated words produced. ● Response is completely off-topic, non-English or unintelligible. ● Poor audio quality means response cannot be scored.
Note	Speech should fully match the descriptor to be awarded the corresponding score. Where a response does not meet all parts of the descriptor, the lower score should be given.				

Contact information

About the British Council
The British Council builds connections, understanding and trust between people in the UK and other countries through arts and culture, education and the English language.

We work in two ways – directly with individuals to transform their lives, and with governments and partners to make a bigger difference for the longer term, creating benefit for millions of people all over the world.

We help young people to gain the skills, confidence and connections they are looking for to realise their potential and to participate in strong and inclusive communities. We support them to learn English, to get a high-quality education and to gain internationally recognised qualifications. Our work in arts and culture stimulates creative expression and exchange and nurtures creative enterprise.

We connect the best of the UK with the world and the best of the world with the UK. These connections lead to an understanding of each other's strengths and of the challenges and values that we share. This builds trust between people in the UK and other nations which endures even when official relations may be strained.

We work on the ground in more than 100 countries. In 2019–20, we connected with 80 million people directly and with 791 million overall, including online and through our broadcasts and publications.

Contact EnglishScore
For questions about the test, including content development, test scoring, security or certification, please contact:

EnglishScore
Scale Space
58 Wood Lane
London W12 7RZ
United Kingdom
contact@englishscore.com

References

Cambridge University Press, 2015. *Reference Level Descriptions* [Online]. Available from: http://englishprofile.org/the-cefr/reference-level-descriptions [Accessed 24 November 2022].

Council of Europe, 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment* [Online]. Strasbourg: Council of Europe. Available from: rm.coe.int/1680459f97 [Accessed 24 November 2022].

Linacre, J. M., 2012. *Winsteps Rasch Tutorial 2* [Online]. Available from: https://www.winsteps.com/a/winsteps-tutorial-2.pdf [Accessed 20 March 2023].

Weir, C. J., 2005. *Language Testing and Validation: An Evidence-Based Approach*. Hampshire: Palgrave MacMillan.

BRITISH COUNCIL

EnglishScore