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.

March 2025

Table of contents

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 by the British Council (www.britishcouncil.org), the United Kingdom's international organisation for cultural relations and educational opportunities.

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 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

[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.

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

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:

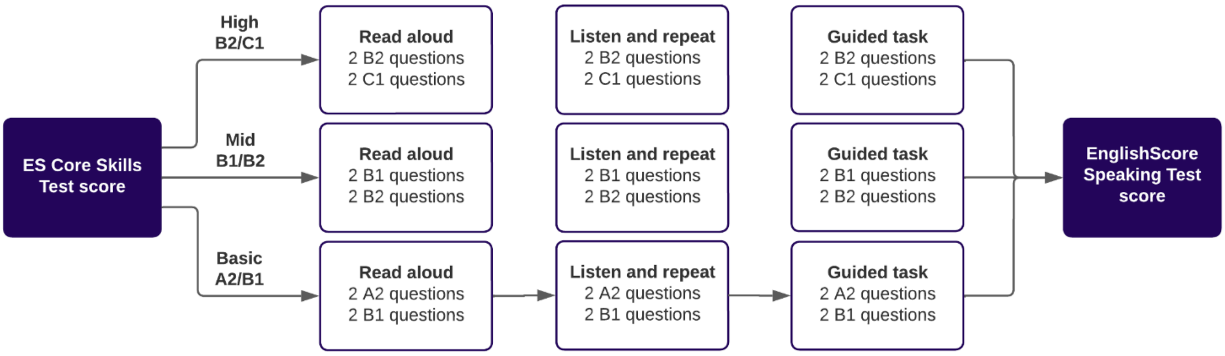
Basic (Breakthrough +):	CEFR levels A2 and B1
Mid (Threshold +):	CEFR levels B1 and B2
High (Vantage +):	CEFR levels B2 and C1

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 three item types:

- Read Aloud
- Listen and Repeat
- Guided Task

Figure 1. *The EnglishScore Speaking Test structure.*



b. Domains assessed

The speaking assessment tests spoken English in the **personal**, **public** and *generic* **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.*

III. The content of EnglishScore

a. Overview

Assessment structure

Each of the three levels (A2/B1, B1/B2, B2/C1) in the assessment contains the following item types:

- Read aloud
- Listen and repeat
- Guided task

Table 1. *The EnglishScore Speaking Test: overview of the test sections.*

Item type	CEFR level	Example focus	Input length and level	Domain	Nature of info, topic familiarity
Read aloud	**A2**	Read out phrase on screen	~10 words	Personal and public	Concrete, familiar
	B1	Read out phrase on screen	~14 words	Personal and public	Concrete, familiar
	B2	Read out phrase on screen	~18 words	Public and professional	Mix of concrete and abstract, mix of familiar and unfamiliar
	C1	Read out phrase on screen	~22 words	Public and professional	Abstract, unfamiliar
Listen and repeat	**A2**	Listen to speaker, repeat heard phrase	~6 words	Personal and public	Concrete, familiar
	B1	Listen to speaker, repeat heard phrase	~9 words	Personal and public	Concrete, familiar
	B2	Listen to speaker, repeat heard phrase	~12 words	Public and professional	Mix of concrete and abstract, mix of familiar and unfamiliar
	C1	Listen to speaker, repeat heard phrase	~14 words	Public and professional	Abstract, unfamiliar
Guided task	**A2**	Respond to a question (accompanied by an image) by giving a(n) opinion, idea, description, explanation	-	Personal and public	Concrete, familiar

		etc.			
	B1	Respond to a question by giving a(n) opinion, idea, description, explanation etc.	-	Personal and public	Concrete, familiar
	B2	Respond to a question (accompanied by an image) by giving a(n) opinion, idea, description, explanation etc.	-	Public and professional	Mix of concrete and abstract, mix of familiar and unfamiliar
	C1	Respond to a question by giving a(n) opinion, idea, description, explanation etc.	-	Public and professional	Abstract, unfamiliar

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.

Read aloud

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

Connection to language use

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

Instructions

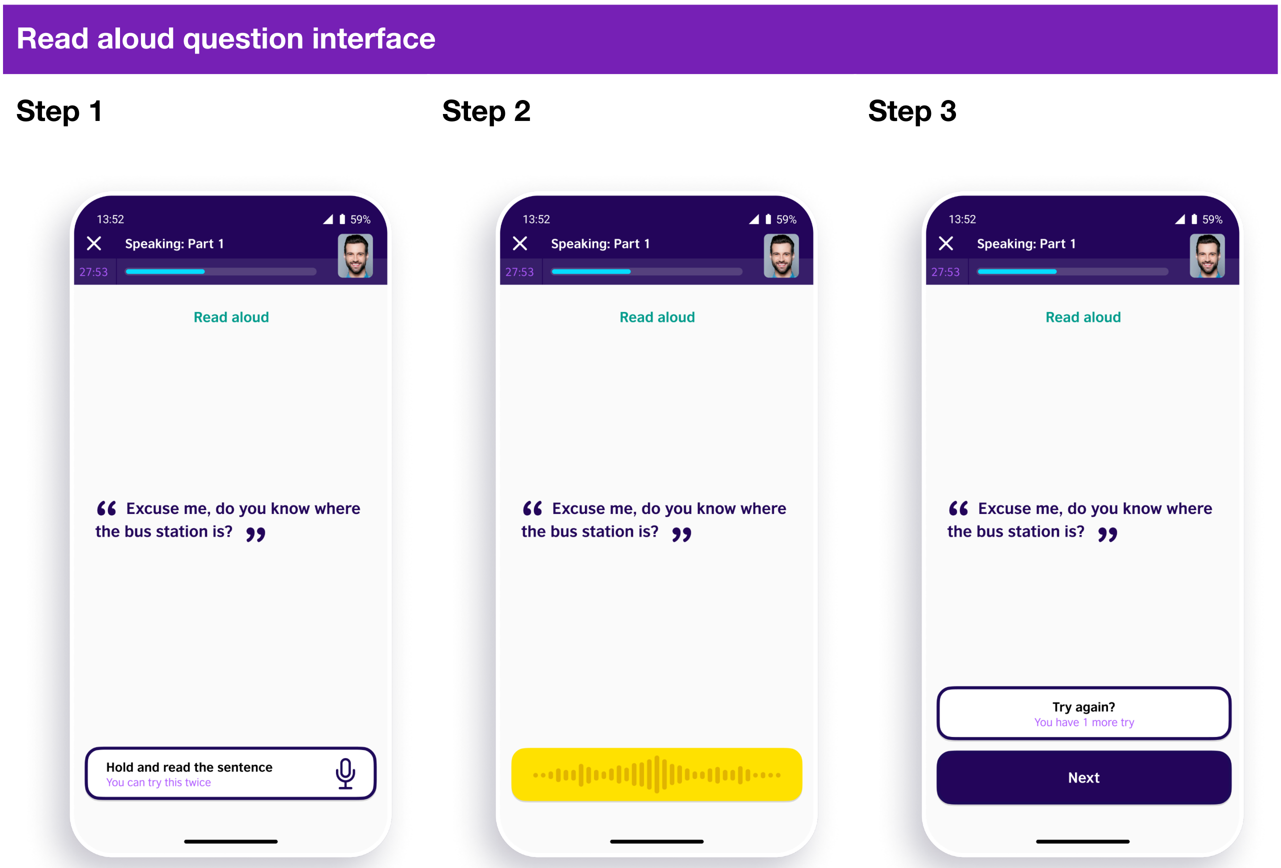
All instructions for the Read aloud item type are given in writing which remains on screen during the assessment. Test-takers can re-record a response if they wish (but cannot listen to their first attempt).

Test 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

Step 2

Step 3

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

Timing

Test-takers are given 45 seconds to complete each question. This time does not change across levels.

The app allows test-takers to see the remaining time at the top of the screen.

What is the input?

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.

What the test-taker needs to do (the expected response)

In Read aloud, test-takers read a sentence out loud and records it. Sentence word count ranges from around 10 to 22 words, as shown in Table 1. The phrase required to be repeated will increase in length and complexity depending on the CEFR level of that item.
The responses will be evaluated based on how accurately test-takers produce the written input. Test-takers must read, understand and produce the sentence, allowing assessment of a test-taker's English pronunciation and fluency proficiency.

[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.

Listen and repeat

Rationale
In this item type, test-takers hear a short sentence and repeat it out loud verbatim. Item word counts range between 5 and 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.

Connection to language use

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.

Instructions

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

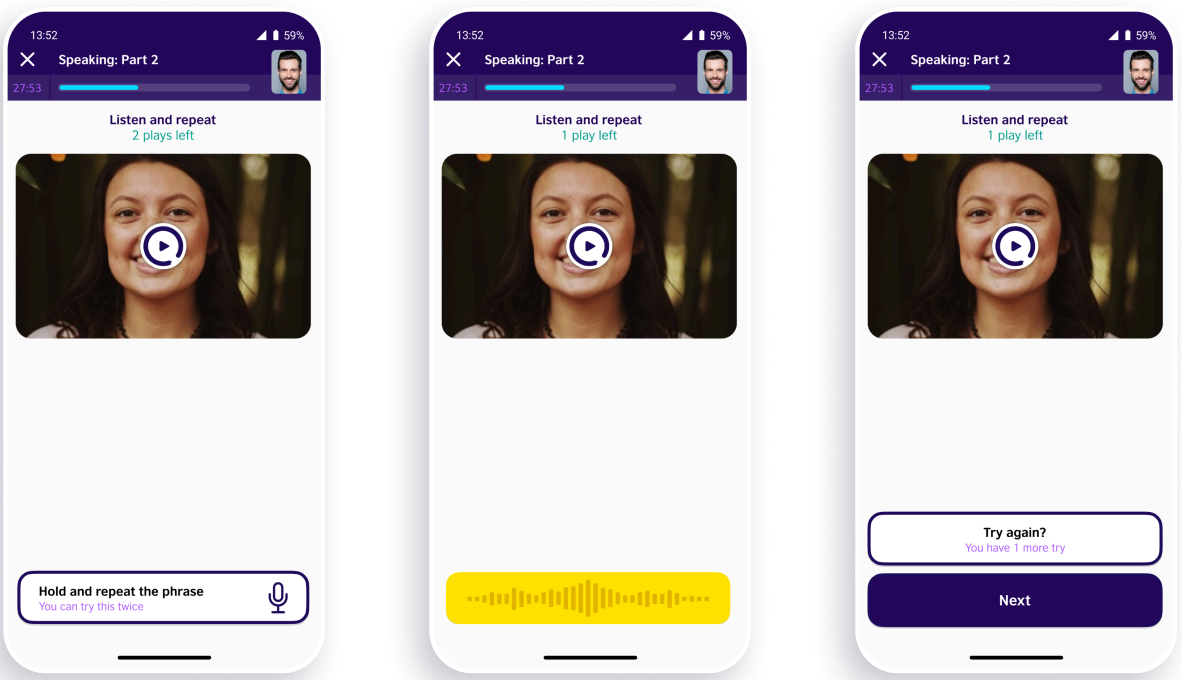
Figure 3 shows the stages of the Listen and repeat question flow for the test-taker.

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

Timing
Test-takers are given 45 seconds to complete each question. This time does not change across levels.

The app allows test-takers to see the remaining time at the top of the screen.

What is the input?

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[4] 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.* 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.

What the test-taker needs to do (the expected response)

In Listen and repeat, test-takers listen to a sentence, repeat it out loud and record it. Sentence word count ranges from around 6 to 14 words, as shown in Table 1. The phrase required to be repeated will increase in length and complexity depending on the CEFR level of that item.
The responses will be evaluated based on how accurately test-takers produce the spoken input. Test-takers must listen to, understand and produce the sentence, allowing assessment of a test-taker's English pronunciation and fluency proficiency.

Guided task

Rationale
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.

Connection to language use

[4] 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.

This task type assesses extended speaking turns, measuring overall communication including providing relevant and supporting detail in responses, coherence and awareness of context.

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

Instructions

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

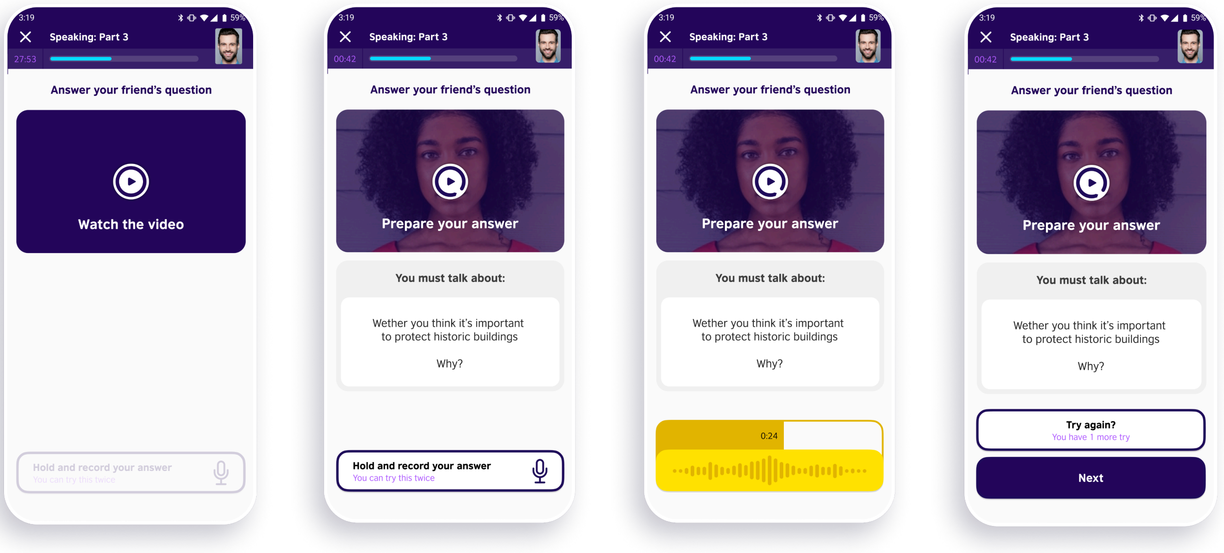
Figures 4 and 5 show the stages of the Guided task question flow for the test-taker.

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

3:19 59%

Speaking: Part 3

27:53

Answer your friend's question

Watch the video

Hold and record your answer
You can try this twice

Step 2

3:19 59%

Speaking: Part 3

00:42

Answer your friend's question

Prepare your answer

You must talk about:

Wether you think it's important to protect historic buildings

Why?

Hold and record your answer
You can try this twice

Step 3

3:19 59%

Speaking: Part 3

00:42

Answer your friend's question

Prepare your answer

You must talk about:

Wether you think it's important to protect historic buildings

Why?

0:24

Step 4

3:19 59%

Speaking: Part 3

00:42

Answer your friend's question

Prepare your answer

You must talk about:

Wether you think it's important to protect historic buildings

Why?

Try again?
You have 1 more try

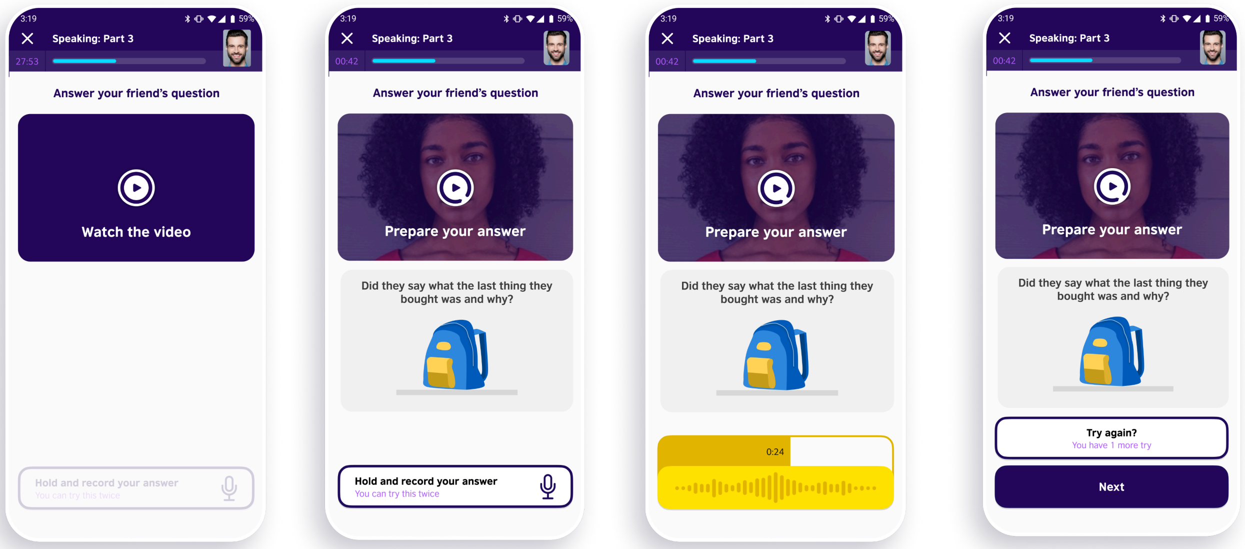
Next

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



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

Timing
Test-takers are given between 100-150 seconds to complete each question. The exact time is level-dependent.

The app allows test-takers to see the remaining time at the top of the screen.

What is the input?

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[5] 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.* 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.

What the test-taker needs to do (the expected response)

In Guided task, test-takers read and/or listen to a task, and record their response to it. The task given to the test taker will increase in length and complexity depending on the CEFR level of that item.
The responses will be evaluated based on how accurately test-takers produce the written input. Test-takers must read and understand the task prompt, and then produce a response, allowing assessment of a test-taker's English spoken proficiency.

b. Writing the assessment material for EnglishScore

Test writer qualifications

EnglishScore writers are teachers of English with a teaching qualification such as a Master's 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.

[5] 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.

Test writer guidance

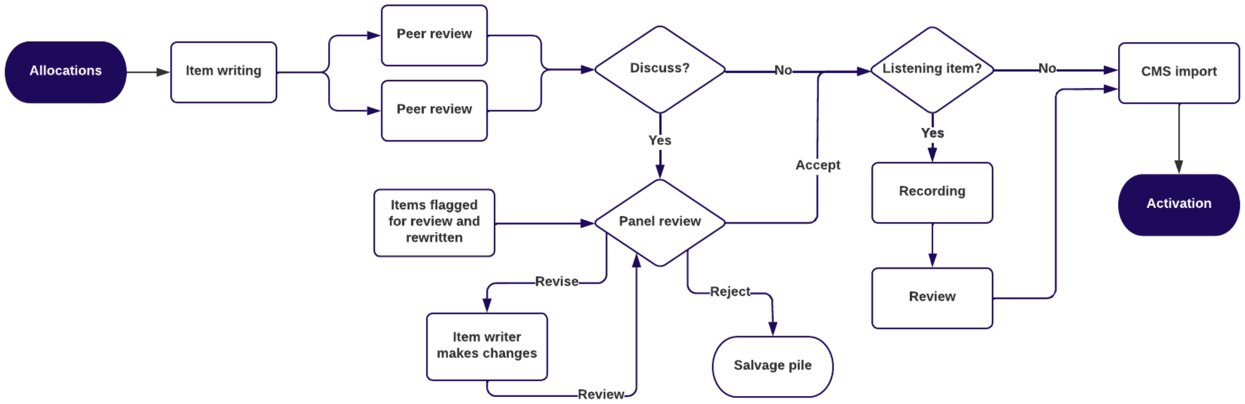
To ensure that the content follows the developers' intentions and that it is parallel across different versions of EnglishScore, when preparing material, the writers follow detailed writer guidelines for each section of the test. These include examples of good (and poor) tasks, wordlists, lists of grammatical 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, regular 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 automated AI scoring that reflects how well a test-taker can communicate with people from a variety of different backgrounds, cultures and English language levels.

Read aloud

This item type is evaluated by automated AI scoring for:

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

Listen and repeat

This item type is evaluated by automated AI scoring for:

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

Guided task

This item type is evaluated by automated AI scoring for:

- Communication
    - Task attempt
    - Response coherence
    - Context awareness

b. Score reporting

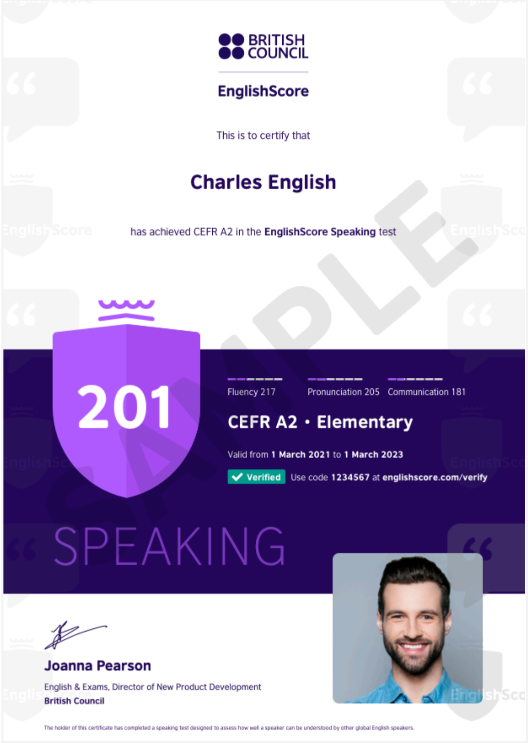
On completing the EnglishScore Speaking Test, the test-taker is provided with an onscreen report stating their overall ‘EnglishScore’ speaking score, with separate score breakdowns for pronunciation, fluency and communication.

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

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 the test.

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 uses automated AI scoring to generate 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 over 8000 test-taker responses. 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 scores to arrive at a model that correlates strongly with experts. The current model has a strong correlation of 0.87.

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 – can the test-taker produce an answer that is relevant to the prompt and contains additional detail and supplementary information where appropriate? Factors such as coherence, development of ideas, and awareness of context lead to a higher score in this subskill.

d. 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.

This report will be periodically updated with additional test and item performance metrics.

Item performance

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 2 features or traits.

The task types were:

Listen and repeat

Read aloud

These task types were scored using AI 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.

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.

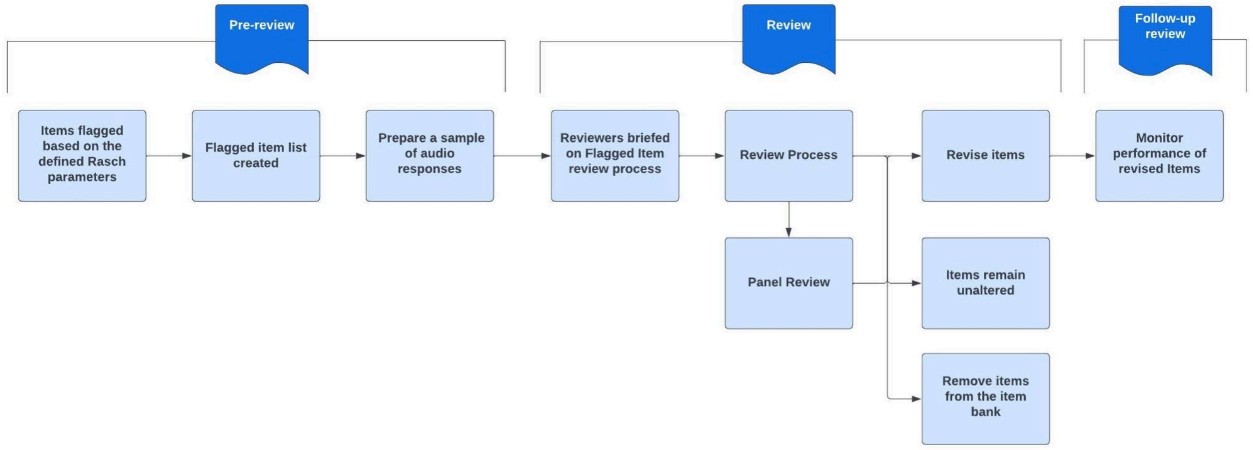
Review stage:

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
Revise items
Monitor performance of revised Items
Panel Review
Items remain unaltered
Remove items from the item bank

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 us at 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.