BRITISH
COUNCIL

EnglishScore

Writing Test Purpose and Content

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

July 2023

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 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 their workforce or assess a future employee's language ability, by universities and language

[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.

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. Users 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.

Writing Test

The EnglishScore writing assessment complements and supports the Core Skills Test and is designed to measure a test-taker’s writing 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

Development of EnglishScore

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 Writing 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 written performance, and these, along with other factors such as test-taker characteristics, are considered when designing test tasks.

The EnglishScore Writing Test is a single test, assessing a test-taker's writing proficiency. The difficulty of the test items included 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 abilities. 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.

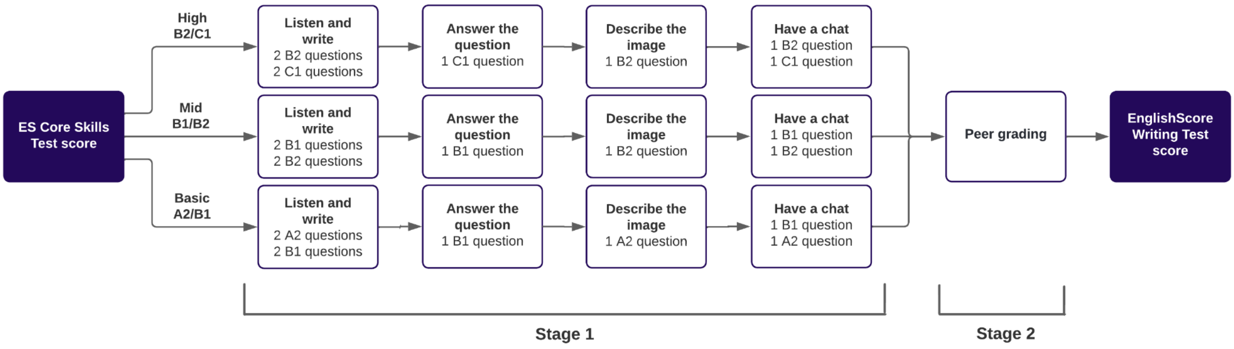
The Writing Test consists of two main stages.

Stage 1 includes these four sections:

- Listen and write
- Answer the question
- Describe the image
- Have a chat.

Stage 2, Peer grading, requires test-takers to rate other test-takers' responses.

Figure 1. *The EnglishScore Writing Test stages.*


ES Core Skills Test score
High B2/C1
Mid B1/B2
Basic A2/B1
Listen and write 2 B2 questions 2 C1 questions
Answer the question 1 C1 question
Describe the image 1 B2 question
Have a chat 1 B2 question 1 C1 question
Listen and write 2 B1 questions 2 B2 questions
Answer the question 1 B1 question
Describe the image 1 B2 question
Have a chat 1 B1 question 1 B2 question
Listen and write 2 A2 questions 2 B1 questions
Answer the question 1 B1 question
Describe the image 1 A2 question
Have a chat 1 B1 question 1 A2 question
Peer grading
EnglishScore Writing Test score
Stage 1
Stage 2

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

b. Domains assessed

The Writing Test assesses writing skills required in the **personal**, **public** and *generic* **occupational** domains, with an emphasis on workplace English. 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, using public transport and interacting 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.*

To understand more about the needs of the test-taker population, a short questionnaire was designed. The aim of this questionnaire was to gain some information about test-taker needs, which can be useful to understand the domain needs. In total, 186 people answered the questionnaire from different countries and contexts. The respondents were given a series of questions about the task types that they use in their workplaces. The most common mode of communication was instant messaging, such as Microsoft Teams, Slack and WhatsApp. Respondents also reported that they sometimes use email and very occasionally write reports and proposals or meeting agendas and minutes, or marketing and social media posts. Respondents reported that the most common functions when writing were giving someone information (82%), asking someone for information (76%) and giving one’s opinion about something (71%).

These answers suggest that being able to communicate with work colleagues as well as people outside their workplaces is important for people who want to achieve English proficiency for general workplace purposes.

III. The content of EnglishScore

a. Overview

Stage 1 – Writing assessment
This section covers Stage 1, the writing assessment stage. Each of the three levels (A2/B1, B1/B2, B2/C1) in Stage 1 contains the following sections or parts:

- Listen and write
- Answer the question
- Describe the image
- Have a chat.

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

Test Sections	CEFR level	Example focus	Input length and level	Expected response length	Domain	Nature of info, topic familiarity
Listen and write (Dictation)	**A2**	Listen to a sentence and type	Around 5–9 words	-	Personal and public	Concrete, familiar
	B1	Listen to a sentence and type	Around 8–10 words	-	Personal and public	Concrete, familiar
	B2	Listen to a sentence and type	Around 10–13 words	-	Public and profession al	Mix of concrete and abstract, mix of familiar and unfamiliar
	C1	Listen to a sentence and type	Around–15 words	-	Public and profession al	Abstract, unfamiliar
Answer the question	**A2**	Read a question and type your answer	-	20–40 words	Personal and public	Concrete, familiar
	B1	Read a question and type your answer	-	20–40 words	Personal and public	Concrete, familiar
	B2	Read a question and type your	-	50–70 words	Public and profession	Mix of concrete and abstract,

Test Sections	CEFR level	Example focus	Input length and level	Expected response length	Domain	Nature of info, topic familiarity
		answer			al	mix of familiar and unfamiliar
	C1	Read a statement and type your answer	-	50–70 words	Public and profession al	Abstract, unfamiliar
Describe the image	**A2**	Look at an image and describe it	-	20–50 words	Personal and public	Concrete, familiar
	B1	Look at an image and describe it	-	20–50 words	Personal and public	Concrete, familiar
	B2	Look at an image and describe it	-	30–50 words	Public and profession al	Mix of concrete and abstract, mix of familiar and unfamiliar
	C1	Look at an image and describe it	-	30–50 words	Public and profession al	Abstract, unfamiliar
Have a chat	**A2**	Read the prompt and chat with the person	Between 10 and 40 words, depending on the stages	15–30 words	Personal and public	Concrete, familiar
	B1	Read the prompt and chat with the person	Between 10 and 40 words, depending on the stages	15–40 words	Personal and public	Concrete, familiar
	B2	Read the prompt and chat with the person	Between 10 and 40 words, depending on the stages	30–50 words	Public and profession al	Mix of concrete and abstract, mix of familiar and unfamiliar
	C1	Read the prompt and chat with the person	Between 10 and 40 words, depending on the stages	30–50 words	Public and profession al	Abstract, unfamiliar

Listen and write

Rationale

In Listen and write (or Dictation), test-takers hear a short sentence and write it verbatim. Test-takers can listen to the sentence twice. The written phrase or statement expected from test-takers will increase in length and complexity depending on the CEFR level of that item.

Connection to language use

This section is similar to writing activities that are commonly encountered in personal, public and general occupational settings around the world. It is important to understand the spoken input and write it down. Such activities may be useful in contexts, for instance, where a person listens to a customer/colleague and writes down the input.

Instructions

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

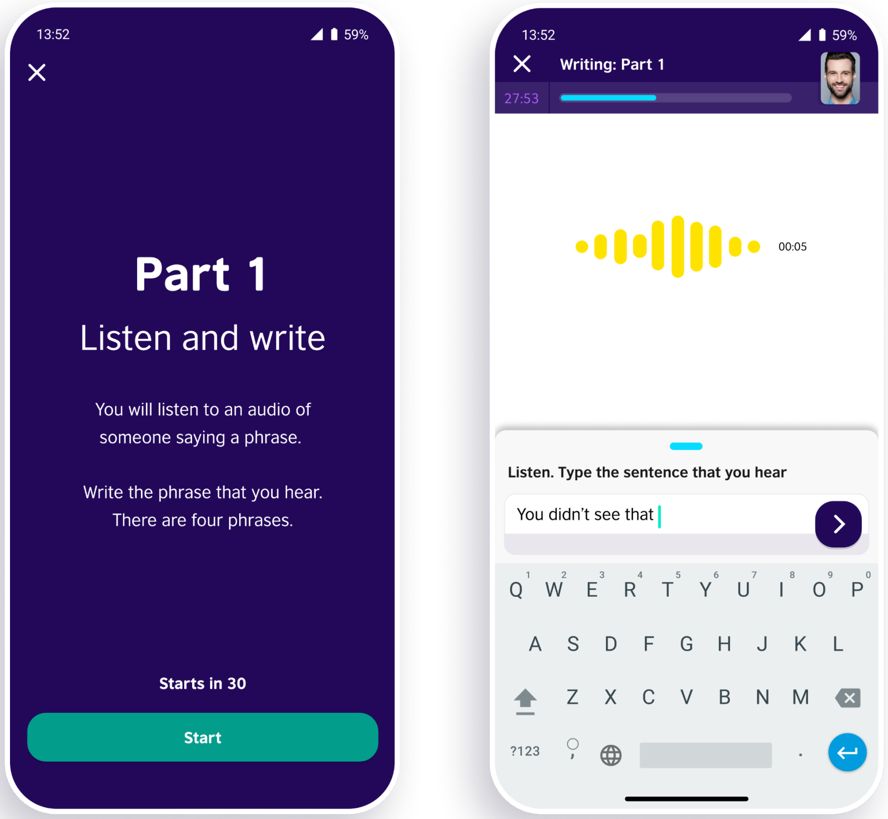
Test steps:
Figure 2 shows the steps of Listen and write for the test-taker.

> **Step 1:** First, the test-taker is instructed to listen to the spoken input and then write (type in) their responses.
> **Step 2:** The test-taker types their responses.
> **Step 3:** The test-taker has the option to listen to the input twice. They are not allowed to listen to the spoken input more than twice. The test-taker is able to see their typed text.

The app does not allow the test-taker to copy and paste from a different source, nor does it correct grammar mistakes or allow for grammar-checking tools to operate while the test-taker types their answer.

The app does not allow for any autosuggestions or corrections to the language.

Figure 2. *Listen and write.*



You will listen to an audio of someone saying a phrase.

Write the phrase that you hear. There are four phrases.

Listen. Type the sentence that you hear

You didn't see that

Timing

Test-takers are given ninety 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, weather and others.

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. The items are approved by the EnglishScore Senior Assessment Manager.

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 two-chat instructions. For the 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.

The input is prepared by item writers specially for EnglishScore. Items are written to reflect the spoken features, grammatical structures and vocabulary test-takers would encounter in the relevant domains. Writers use English Profile (www.englishprofile.org) Reference Level Descriptions for English to guide the choice of vocabulary and grammatical structures used in the item.

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

In Listen and write, test-takers hear a sentence and type it. Sentence word count ranges from around 5 to 15 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. A current word count is displayed to the test-taker as they type.

The responses will be evaluated based on how accurately test-takers write the heard input. Test-takers must listen to, understand and produce the sentence, allowing assessment of a test-taker's knowledge of written language (grammar, vocabulary, syntax and spelling) in English.

[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.

Answer the question

Rationale

In this section, test-takers read a short prompt about a context and write their reply. Expected response word counts range between 20 and 70 words. The prompt and expected response increase in length and complexity depending on the CEFR level of that item. The prompt remains on screen during the test.

Connection to language use

This section is similar to writing activities that are commonly encountered in personal, public and general occupational settings around the world. As the needs analysis showed, such activities may be useful in contexts where people are asked to briefly write their opinions about a problem, statement or issue and give their suggestions or advice.

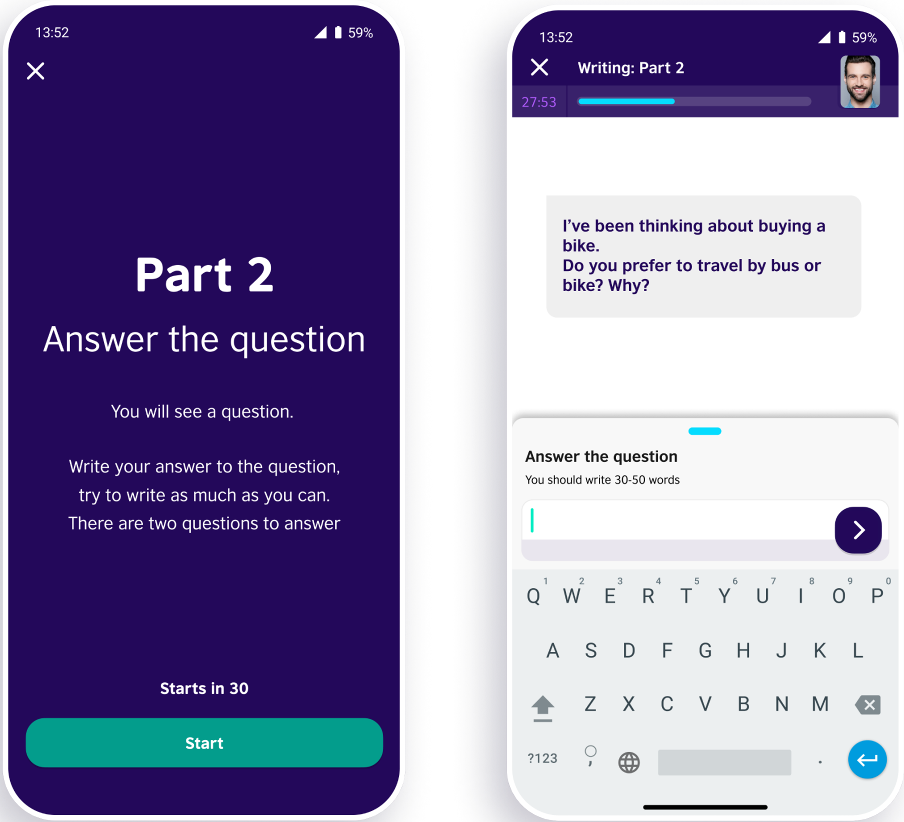
Instructions

The task instructions are given in English. Figure 3 shows the steps of Answer the question for the test-taker.
Step 1: First, the test-taker is instructed to read the statement and then write (type in) their responses.
Step 2: The test-taker types their responses. The test-taker is able to see their typed text.

The app does not allow the test-taker to copy and paste from a different source, nor does it correct grammar mistakes or allow for grammar-checking tools to operate while the test-taker types their answer.

The app does not allow for any autosuggestions or corrections to the language.

Figure 3. Answer the question.



Part 2

Answer the question

You will see a question.

Write your answer to the question, try to write as much as you can. There are two questions to answer

I've been thinking about buying a bike.
Do you prefer to travel by bus or bike? Why?

Answer the question

You should write 30-50 words

Timing

Test-takers are given seven minutes to complete this section. 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* and others.

Sources of input

Item content is prepared by item writers specially for EnglishScore. Items are written to reflect the text 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. The items are approved by the EnglishScore Senior Assessment Manager.

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.*

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.

The input is prepared by item writers specially for EnglishScore. Items are written to reflect the grammatical structures and vocabulary test-takers would encounter in the relevant domains. Writers use English Profile (www.englishprofile.org) Reference Level Descriptions for English to guide the choice of vocabulary and grammatical structures used in the item.

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

In Answer the question, test-takers read a statement about a context and type their answer. The question that needs to be answered will increase in length and complexity dependent on the CEFR level of that item.

The length of the expected response varies between 20 and 40 words at A2/B1 levels and 50 and 70 words at B2/C1 levels. A current word count is displayed to the test-taker as they type.

At A2/B1 levels, the functions to be elicited from the items are mainly *opinion* and *concrete* information. At B2/C1 levels, the functions of the expected response are mainly *abstract/hypothetical*, which include *opinion/speculation/problem and solution/recommendation*.

The responses will be evaluated based on grammar accuracy and range, spelling, the use of vocabulary, organisation of ideas and the content (task completion).

[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.

Describe the image

Rationale

In Describe the image, test-takers see an image and are asked to describe it. The written response expected from test-takers will increase in length and complexity dependent on the CEFR level of that item.

Connection to language use

The tasks are similar to writing activities that are commonly encountered in personal, public and general occupational settings around the world. Such activities may be useful in the contexts where a person describes scenarios and speculates the context in workplace settings.

Instructions

The task instructions are given in English.

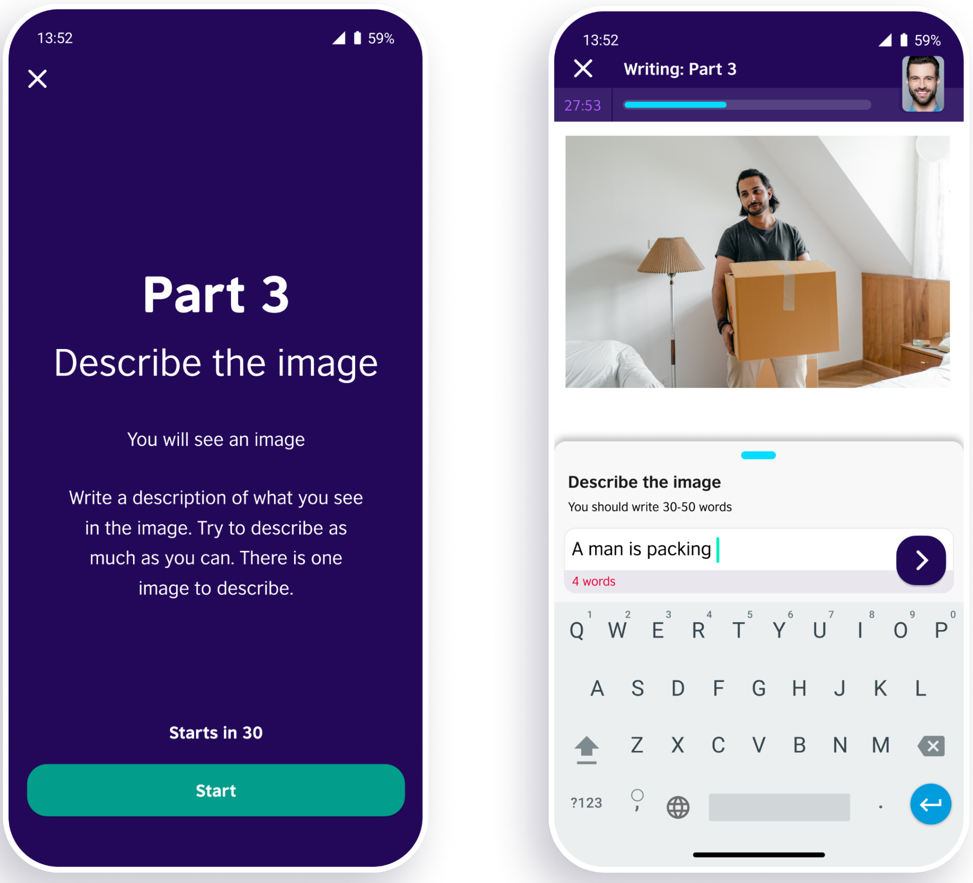
Test steps:
Figure 4 shows the steps of Describe the image for the test-taker.

Step 1: First, the test-taker is instructed to look at the image and then write (type in) their responses.

Step 2: The test-taker types their responses.

The app does not allow the test-taker to copy and paste from a different source, nor does it correct grammar mistakes or allow for grammar-checking tools to operate while the test-taker types their answer.
The app does not allow for any autosuggestions or corrections for the language.

Figure 4. *Describe the image.*



13:52 59%

Part 3

Describe the image

You will see an image

Write a description of what you see in the image. Try to describe as much as you can. There is one image to describe.

Starts in 30

Start

13:52 59%

Writing: Part 3

27:53

Describe the image

You should write 30-50 words

A man is packing

4 words

Timing
Test-takers are given five minutes to complete this section. This time does not change across levels.

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

What is the input?

The input
Images in the input do not contain excessive details that may not be visible on a small screen. These images avoid contexts that may advantage or disadvantage some test-takers in terms of their background knowledge or topic familiarity or that may not be appropriate for some cultures.

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* and others.

Sources of input

Item content is prepared by item writers specially for EnglishScore. The images are written to reflect the communication features test-takers would encounter in the relevant domains. Item writers are trained on writing this item including selecting appropriate and suitable images and writing suitable prompts. The relevance and suitability of the images and prompt to elicit the expected response are reviewed and approved by the EnglishScore Senior Assessment Manager.

Nature of input

Images in this section of the test relate to the personal, public or professional domains. The same image is used across all levels.

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

In Describe the image, test-takers see an image and are asked to describe it in a written form. The text required from test-takers to be written will increase in length and complexity dependent on the CEFR level of that item. The expected response length is between 20 and 50 words at A2/B1 levels and 30 and 50 words at B2/C1 levels. A current word count is displayed to the test-taker as they type.

At A2/B1 levels, the function to be elicited from the item is mainly *description*. At B2/C1 levels, the functions of the expected response are mainly *description/speculation/inference/interpretation*.

The responses will be evaluated based on how the test-takers achieve the task completion. Test-takers must describe the image and interpret it, allowing assessment of a test-taker's knowledge of written language (grammar, vocabulary, syntax and spelling) in English.

Have a chat

Rationale

In the Have a chat section of the test, test-takers read a context or scenario which is often 15 words long. This input remains on the screen while test-takers are allowed to plan and write (type) their answers. Test-takers respond to questions by typing their answers. The response expected from test-takers will increase in length and complexity dependent on the CEFR level of that item.

Connection to language use

The tasks are similar to writing activities that are commonly encountered in personal, public and general occupational settings around the world. For example, customers/users or visitors often contact a company through the website chat. Sometimes, the respondents are automated intelligence (AI)-generated robots but sometimes, they are human. Regardless of the nature of the chatting agent, whether human or robot, people need to type and explain their concerns. Likewise, it is also important for human chat agents to be able to understand customers' concerns and respond to them.

Instructions

The task instructions are given in English.

Test steps:

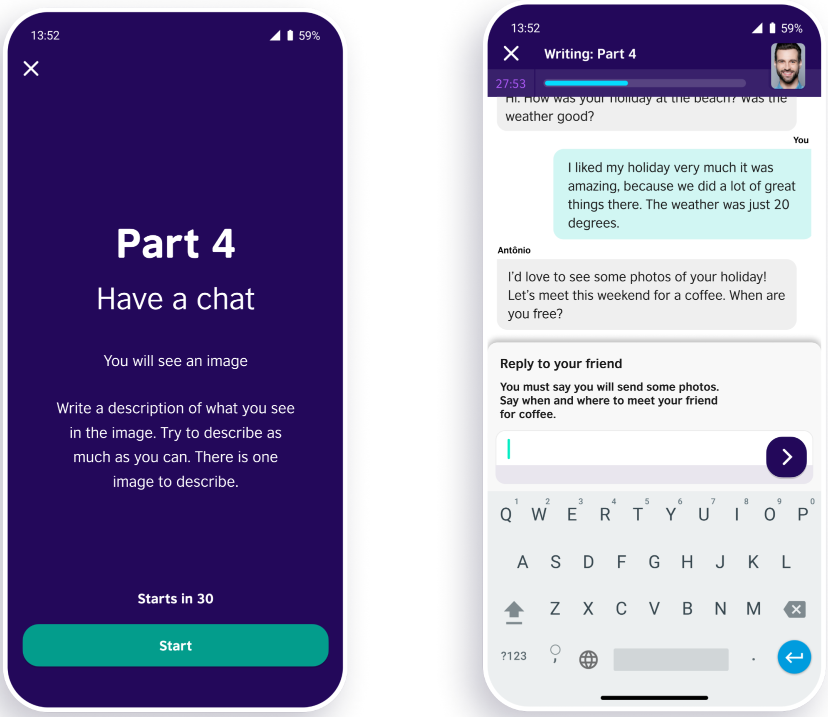
Figure 5 shows the steps of Have a chat for the test-taker.

Step 1: Introduction. First, a topic is introduced to set a context. For instance, '*you are chatting with your friend about your holiday*'. This part may consist of about 15 words.

Step 2: Chat title. This reminds the test-taker of the topic they are discussing. The test-taker sees the chat title, e.g. '*Chat with your friend*'.

Step 3: Response prompt. The test-taker reads the prompt which gives an instruction on what questions they should address or answer in their responses. The prompt facilitates the test-taker response as it outlines what they should address in their responses. The test-taker continues chatting until the prompt finishes.

Figure 5. *Have a chat.*



13:52 59%

Part 4

Have a chat

You will see an image

Write a description of what you see in the image. Try to describe as much as you can. There is one image to describe.

Starts in 30

Start

13:52 59%

Writing: Part 4

27:53

Hi. How was your holiday at the beach? Was the weather good?

You

I liked my holiday very much it was amazing, because we did a lot of great things there. The weather was just 20 degrees.

Antônio

I'd love to see some photos of your holiday! Let's meet this weekend for a coffee. When are you free?

Reply to your friend

You must say you will send some photos. Say when and where to meet your friend for coffee.

Timing

Test-takers are given five minutes to complete this whole section. This time does not change across levels.

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, weather, customer service chats/appointments, company requests* and others[9].

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. The items are approved by the EnglishScore Senior Assessment Manager.

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*.

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.

The input is prepared by item writers specially for EnglishScore. Items are written to reflect the spoken features, grammatical structures and vocabulary test-takers would encounter in the relevant domains. Writers use English Profile (**www.englishprofile.org**) Reference Level Descriptions for English to guide the choice of vocabulary and grammatical structures used in the item.

[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.

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

In the Have a chat section of the test, test-takers are prompted to write their responses in the chat form. The response expected from test-takers will increase in length and complexity dependent on the CEFR level of that item.

The expected response length varies across the levels as well as the chat turns. In the first turn when test-takers respond to the chat question, they are expected to produce between 10 and 20 words across all levels. In the second chat turn, test-takers are expected to write detailed responses that will increase in length according to level. At A2 level, the expected response length is between 15 and 30 words; at B1 level, between 15 and 40 words; and at B2 and C1 levels, between 30 and 50 words. A current word count is displayed to the test-taker as they type.

The responses will be evaluated based on how accurately test-takers write and answer the prompt. Test-takers must read the prompt and type in their answers, allowing assessment of their knowledge of written language (grammar, vocabulary, syntax and spelling) in English.

Stage 2 – Peer grading

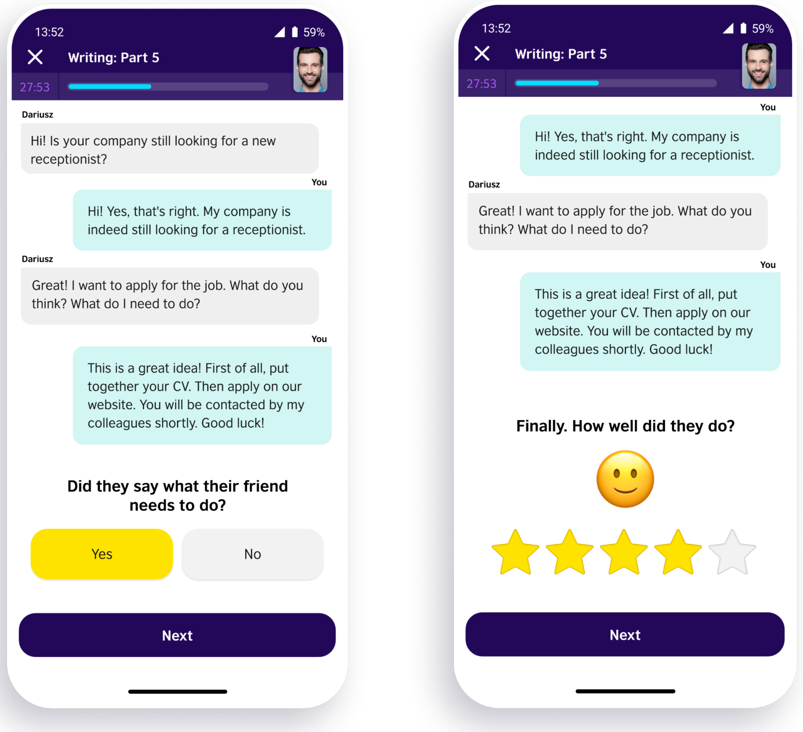
The sections above are in Stage 1 of the EnglishScore Writing Test. Stage 2 of the test is Peer grading, in which peers rate written performances. This stage is compulsory for all test-takers.

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's test is marked by approximately 16 unique peers, all at similar proficiency levels. Sourcing a wide range of peer opinions is believed to reduce bias and ensures a more accurate and reliable measure of a test-taker's proficiency. In addition, EnglishScore continually monitors and reviews peer rater reliability to ensure that peer raters are scoring consistently.

Test-takers 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.

Test-takers peer-mark all the sections of the Writing Test excluding the Listen and write (Dictation) section. The grading questions ask whether the test-taker talked about the prompt and whether they answered the question. The Listen and write section is scored using automated AI scoring.

Figure 6. *Peer grading.*



Dariusz
Hi! Is your company still looking for a new receptionist?

You
Hi! Yes, that's right. My company is indeed still looking for a receptionist.

Dariusz
Great! I want to apply for the job. What do you think? What do I need to do?

You
This is a great idea! First of all, put together your CV. Then apply on our website. You will be contacted by my colleagues shortly. Good luck!

Did they say what their friend needs to do?

You
Hi! Yes, that's right. My company is indeed still looking for a receptionist.

Dariusz
Great! I want to apply for the job. What do you think? What do I need to do?

You
This is a great idea! First of all, put together your CV. Then apply on our website. You will be contacted by my colleagues shortly. Good luck!

Finally. How well did they do?

★★★★☆

Grading steps

Figure 6 shows the stages of peer grading for the test-taker.

Step 1: The rater reads the written responses of other peers.
Step 2: The first grading question asks the rater to choose from a five-star scale for evaluating if test-takers answered the question.
Step 3: The second grading question asks the rater how well the peer did on a five-star scale.

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

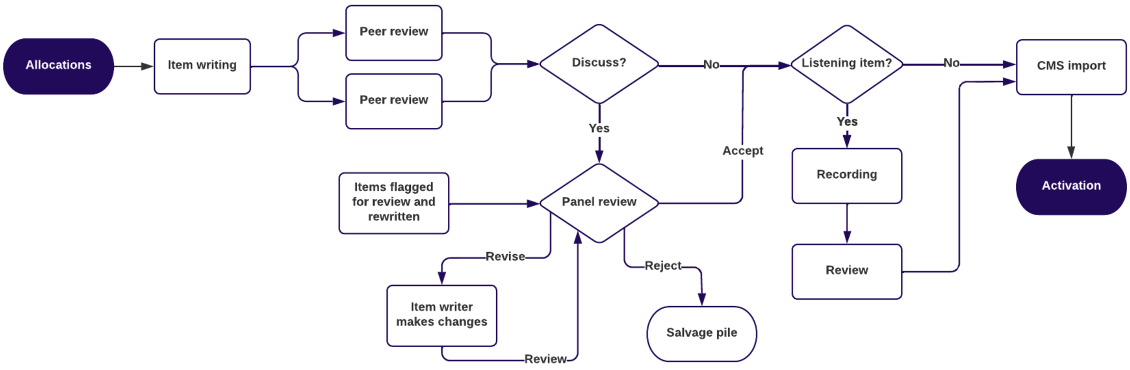
To ensure that the content follows the developers' intentions and to ensure that it is parallel across different versions of EnglishScore, when preparing materials, the writers follow detailed writer guidelines for each section of the test. These include examples of good (and poor) tasks, wordlists, 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 prompts as well as the items 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 7.

Figure 7. *EnglishScore Writing Test 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 by the panel reviewers and the EnglishScore Senior Assessment Manager 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 the EnglishScore test-takers around the world.The test interface is designed to be accessible to colourblind users.

IV. Scoring

a. Scoring

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

Listen and write

This section is evaluated by automated AI scoring for:

- language accuracy including spelling, and range.

Answer the question

This section is evaluated for:

- language – accuracy, including spelling, and range (AI scoring)
- organisation (AI scoring)
- communication and overall comprehensibility (peer grading).

Describe the image

This section is evaluated for:

- language – accuracy, including spelling, and range (AI scoring)
- organisation (AI scoring)
- communication and overall comprehensibility (peer grading).

Have a chat

This section is evaluated for:

- language – accuracy, including spelling, and range (AI scoring)
- organisation (AI scoring)
- communication and overall comprehensibility (peer grading).

b. Score reporting

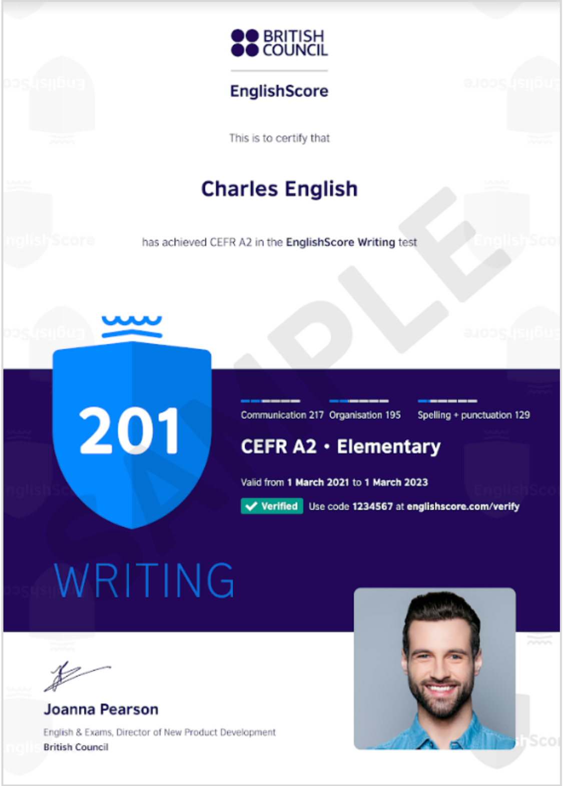
On completing the EnglishScore Writing Test, the test-taker is provided with an onscreen report stating their overall 'EnglishScore' writing score, with separate score breakdowns for skills.

Estimated correspondences to CEFR level 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 writing and the subskills.

This is an example Writing Test certificate:



BRITISH COUNCIL

EnglishScore

This is to certify that

Charles English

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

201

Communication 217 Organisation 195 Spelling + punctuation 129

CEFR A2 • Elementary

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

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

WRITING

Joanna Pearson

English & Exams, Director of New Product Development

British Council

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

Writing 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 writing ability is reported as a writing score from 0 to 599 on the EnglishScore scale, as well as the corresponding CEFR level. In addition, a breakdown of writing subskills (language, organisation, 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 Writing Test scoring model consists of two components: AI scoring and peer grading.

Both data sources feed into the scoring model to give an overall writing score, plus subskill scores in language, organisation and communication.

Scoring model design

A key principle for the scoring model was to ensure alignment with expert raters, i.e. the EnglishScore writing model is designed to score as an expert human rater would. As part of the scoring model development, over 8,000 written 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 building and refining the analytic rating scale descriptors. The rating scale criteria are *language, organisation* and *task response or communication*. 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 writing raters were recruited, trained and certified to use the EnglishScore writing scale descriptors. Each written response was then rated independently by at least two experts. Each expert rater graded eight responses per test. Overall, 8 raters rated 1,000 test-takers' written 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 senior examiner.

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 0.86, and future versions of the scoring model will continue to improve on this.

Writing 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.

Language – can the test-taker produce writing that is accurate and easily understandable to most speakers of the language? Components such as accurate grammar and vocabulary, the use of a range of grammatical structures and vocabulary, and spelling and punctuation lead to a higher score in this subskill.

Organisation – can the test-taker produce writing with coherence and cohesion? Components such as a clearly organised writing with sequence of ideas and use of a range of and appropriate cohesive devices (signposts) 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? Components such as task relevance, appropriate register for the context and developed responses will lead to a higher score in this subskill.

d. Evidence about the score reliability and accuracy

To ensure that the Writing 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 to 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 Writing Test is calculated on a daily basis. The test–retest reliability coefficient for the EnglishScore Writing Test was 0.87, where the same test-taker had taken the test twice in the last 30 days (February – May 2023). The total number of test repetitions used to calculate the coefficient was about 32,000 (N = 31,579).

Item quality and score accuracy

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. The first section discusses the outcomes of the score analysis derived from AI measures, while the second section discusses the outcomes of the scores derived from peer assessment.

AI-scored items

For the analysis, the WINSTEPS Rasch software program was used. Approximately 2,000 test-taker's scores were analysed in each of these 3 levels (easy, mid, hard) with 245 item traits. For the purpose of this analysis, the scores were expressed from 0 to 6.

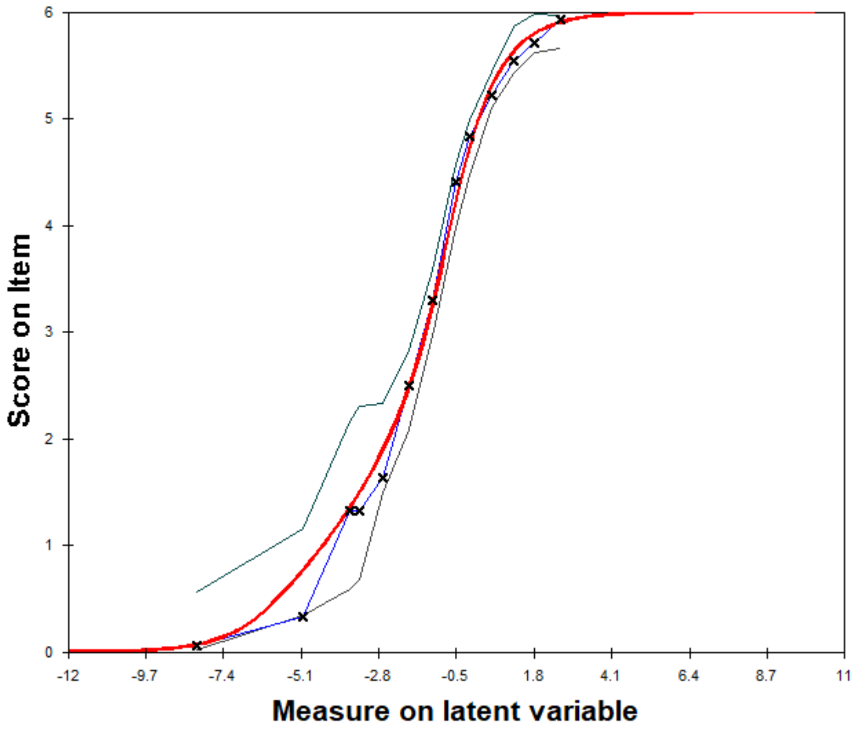
Results

Across the three test levels, **the test reliably discriminated between more and less able test-takers**. The person separation level was around 3 in each level, and the separation reliability was around *0.90*.

In terms of fit, the infit and outfit MnSq values were reviewed to see if the scores meet the Rasch model expectations. By setting the acceptable level between 0.5 and 1.5, we found that around 90% of the sample data fitted the model expectations. The mean square fit statistics index between 0.5 and 1.5 is productive for measurement (Linacre, 2012) for the test's purposes. **This shows that the test items largely aligned with the model expectations.**

To understand the functionality of the items, their empirical and expected item characteristic curve (ICC) plots were analysed.

Figure 8. Example item characteristic curve (ICC).



As Figure 8 shows, the red solid line indicates the model expectation, the blue line shows how the empirical scores match that line or deviate from that line and the grey-green lines show the confidence interval. This is the item with the entry number 67 (infit MnSq = 1.04; outfit MnSq = 1.16). The empirical line is very close to the expected line, and all the scores are within the 95% confidence interval. This is an example for the item fitting the model's expectations.

Items that do not meet the Rasch model expectations are flagged. These flagged items are carefully reviewed by our team of assessment experts. The workflow of the flagged item review process is described in Figure 10.

Peer-scored items

Peer grading reliability

This section reports on the steps taken to ensure the reliability of peer grading.

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 written 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. Required responses are also straightforward and easily understood.

- A minimum of 16 peer gradings are collected to derive a test-taker's score. In contrast to many human-rated writing 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 cannot 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 '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 those of the group. Implementing this step prior to the score calculation ensures that the outlier peer raters' markings do not affect the score accuracy.

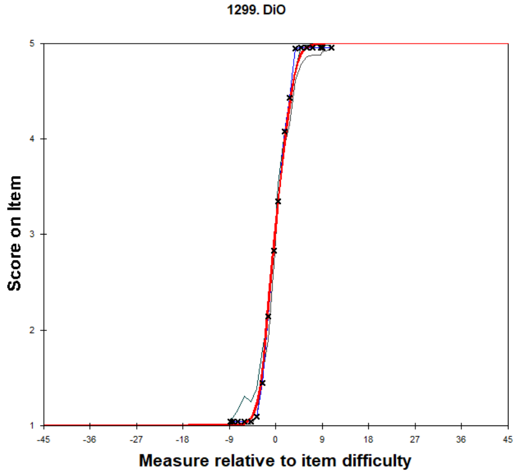
Score analysis

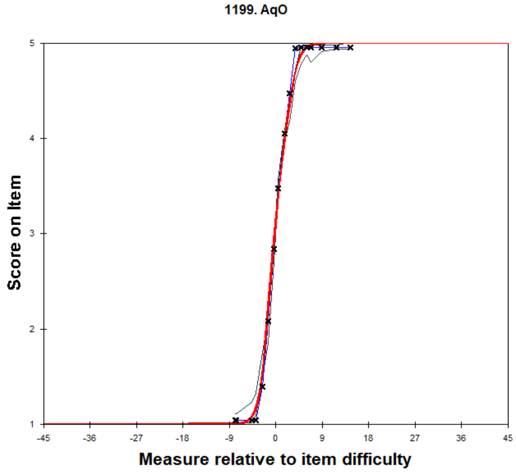
As part of ongoing quality control and review, the scores awarded by peers are also analysed and monitored. For the analysis, the FACETS many-facet Rasch measurement (MFRM) software program was used. The analysis allows us to understand the rater behaviours, the functionality and effectiveness of the score categories and the accuracy of the scores.

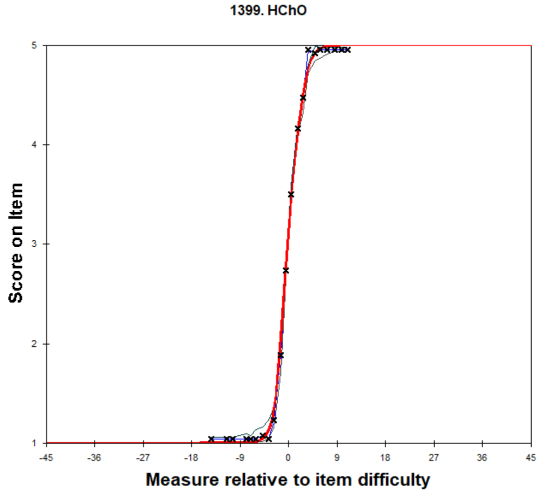
Results

One of the important elements of the Rasch analysis is to understand how the scores align with the expected model. The most recent analysis based on 22,771 test sittings with 491 unique items demonstrated that the tasks fitted the model expectations. For example, as Figure 9 shows, the red solid line indicates the model expectation, the blue line shows how the empirical scores match that line or deviate from that line and the grey-green lines show the confidence interval. These are examples of the Describe image, Answer the question, and Have a chat item types. The empirical line is very close to the expected line, and all the scores are within the 95% confidence interval. This is an example of writing items fitting the model's expectations.

Figure 9. Example ICC.







As part of the ongoing quality control, the outcomes obtained from the Rasch analysis are used to identify problematic items and review those items. The analysis is also useful to identify the peer rater behaviour patterns and feed the outcomes into the scoring algorithms to improve the accuracy of the scores.

Item review and health

The item review process helps to ensure that 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-takers’ writing samples help 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 writing responses.

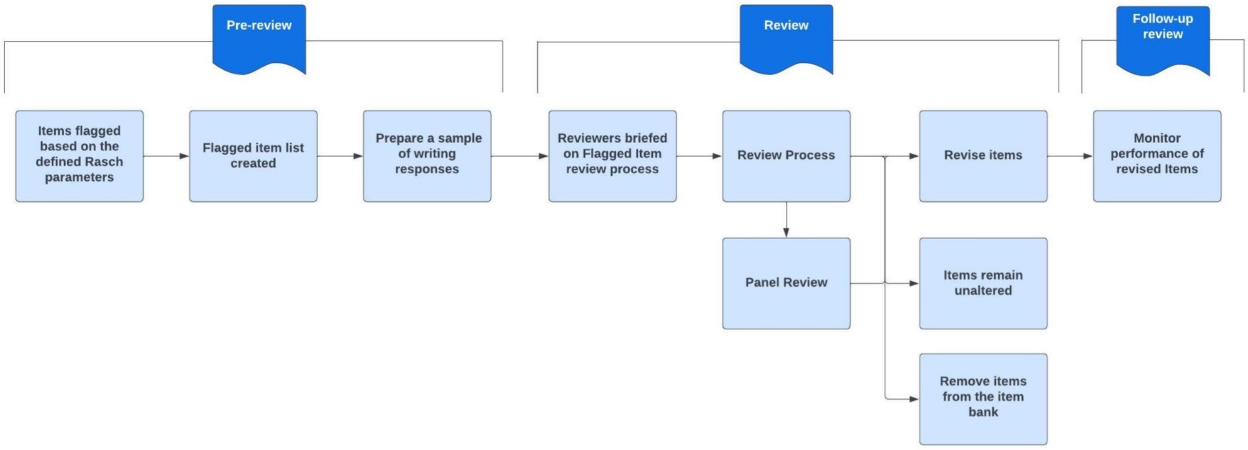
Review stage:

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 writing 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 writing 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 Writing Test.

Appendix 1. Rating Scale for EnglishScore Writing

EnglishScore Writing Descriptors v1			
	Language range and accuracy - grammatical range and accuracy - lexical range - spelling and punctuation	**Organisation** - coherence - cohesion	**Task response** - task attempt - relevance - awareness of context and audience
6 **PROFICIENT**	● Uses a wide range of simple and complex grammatical structures, appropriately and consistently. ● Any errors are minor slips and non-systematic. ● Uses a wide range of vocabulary appropriately and consistently, including idiomatic and technical language. ● Punctuation and spelling are controlled throughout.	● The response is coherent throughout and clearly organised with an appropriate sequence of ideas. ● Uses a wide range of simple and complex cohesive devices appropriately throughout.	● Attempts all parts of the task. ● Responses are fully relevant and appropriately extended and developed. ● Register is consistently appropriate for the genre and context.
5 **ADVANCED**	● A mix of simple and complex structures are generally used appropriately. ● Most of the response is error-free. Where errors do occur, they tend to be slips and do not impact understanding. ● Uses a wide range of vocabulary appropriately, including idiomatic language, with some minor errors. ● Punctuation and spelling are generally controlled throughout. Any errors tend to be slips.	● The response is generally coherent and organised with an appropriate sequence of ideas. ● Uses a mix of simple and complex linkers, with greatest accuracy achieved on the simple ones	● Attempts all parts of the task. ● Responses are generally relevant and appropriately developed. ● Register is mostly appropriate for the genre and context.
4 **GOOD**	● Simple and some complex grammatical structures are used appropriately. ● Where systematic errors occur, they do not cause confusion. ● Uses a range of vocabulary appropriately. Complex vocabulary is sometimes used but may not always be accurate or precise. ● Punctuation and spelling are well	● The response is generally coherent, but it can become difficult to follow complex and/or lengthy explanations. ● Ideas are linked together appropriately but may use a limited range of linkers and/or be repetitive.	● Attempts all parts of the task. ● Most of the responses are relevant and appropriately developed, though there may be some ambiguity/minor repetition. ● Attempts to use appropriate register but not always consistent for the genre and context.
3 **INTERMEDIATE**	● Simple grammatical structures are well controlled and generally error-free. When complex structures are used, they tend to be incorrect and cause confusion. ● Simple vocabulary is used appropriately, but use of complex vocab may be limited, repetitive and/or cause confusion. ● Punctuation and spelling is mostly accurate, and any errors do not impact understanding.	● Simple, short responses are coherent, but longer, more complex ones tend to be incoherent. ● Uses basic linkers to connect ideas, but these may be inconsistently used and/or repetitive/mechanical.	● Attempts most of the task, though some of the minor aspects may not be attempted ● Responses are mostly relevant and developed to some extent, though they sometimes can be repetitive or ambiguous ● Register is generally not appropriate for the genre and context
2 **BASIC**	● Simple grammatical structures are used throughout. Errors are noticeable and can impact understanding. ● Vocabulary choices are simple, repetitive and/or imprecise, and may cause confusion. ● Limited control of spelling and punctuation which may impact comprehension.	● The response is not very coherent, with little progression of ideas ● Very basic linkers may be used to connect ideas, but not always appropriately	● Attempts some tasks but in a limited way. ● Responses may be repetitive, irrelevant or not developed beyond simple explanations. ● Little evidence of awareness of register.
1 **LIMITED**	● Most sentences contain errors which severely impact understanding. ● Simple vocab is used, often inappropriately, which causes confusion. ● Spelling and punctuation errors are frequent and impact comprehension.	● The response is not coherent and appears to be separate, unrelated ideas. ● Very basic linkers may be used, but they do not indicate a logical relationship between ideas.	● 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. ● No evidence of awareness of register.
0 **NO EVIDENCE**	● No language, or only a few isolated words produced. ● Response is completely off-topic, non-English or unintelligible. ● The response is a copy of the prompt and/or instructions.		
Note	Text should 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 23 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 23 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.