OXFORD
TEST OF ENGLISH

OXFORD
UNIVERSITY PRESS

Test Specifications

Test development and validation 2018

Contents

1 Introduction

This paper provides an overview of the development and validation of the Oxford Test of English. It sets out the rationale behind the need for the test, how it was developed, and the procedures employed to ensure and maintain its quality. The development stages include:

- the rationale behind developing the test
- the test design process
- the development of the test specifications
- the procedures for the production of test material
- the processes involved in aligning the test to the Common European Framework of Reference for Languages (CEFR).

2 Test description and rationale

Most educational institutions need a valid and reliable means of assessing students at key stages of their language development – especially in relation to the widely understood levels of the Council of Europe's Common European Framework of Reference for Languages (CEFR). The Oxford Test of English was developed to meet this need for learners of English studying on courses in a wide range of institutions, such as language schools, colleges and universities or company language training programmes. The test content is designed to be suitable for students aged 16 and above.

The starting point for the development of any new test is the perceived needs of the prospective stakeholders, for example the learners, their teachers, institutions and other involved parties, such as educational bodies and employers. Bachman and Palmer, in *Language Assessment in Practice* (Bachman and Palmer, 2010), stress the need to identify and describe the benefits a test can bring to the learners and other key stakeholders. With this in mind, the Oxford Test of English was designed to meet both institutional and individual needs. Many institutions require information on their students' language proficiency, especially at the end of their courses. They need to know whether students are ready to move on to follow higher-level language courses, or pursue further studies or activities that require a specific level of English proficiency. The test also serves the individual learner's need for external verification of their language proficiency for study or career progression.

The Oxford Test of English has been designed to measure language proficiency at CEFR levels B2, B1 and A2. Performance below level A2 is indicated as 'Below A2' in test results.

The content of the test is independent of any specific course of study, and reflects a wide range of English language learning programmes. It is therefore ideally suited for measuring students' general proficiency in English at key points in their learning programmes.

The Oxford Test of English focusses on English language learners' ability to both understand and communicate in English, as measured by four modules:

- Speaking
- Listening
- Reading
- Writing.

All modules are delivered entirely online and can be taken individually, or in any combination, on an on-demand basis.

3 Quality assurance

The Oxford Test of English is produced by Oxford University Press (OUP), a department of the University of Oxford. As a result of quality audits carried out by the University's Department of Continuing Education on behalf of the University of Oxford Education Committee, the University of Oxford officially certifies the Oxford Test of English.

The audits represent a continuous process aimed at maintaining and improving the quality of the Oxford Test of English. They involve scrutiny of the different stages of design, production, and administration. The process continues beyond the launch of the test and includes regular reviews of test administrations to ensure that every test taker receives a fair and valid result.

4 The test development process

The test was developed through an iterative design process (see Figure 1), involving:

- initial test design
- drafting of specifications
- production of sample materials
- reviews by internal assessment staff and external assessment consultants
- modification on the basis of the reviews
- trialling with students in teaching centres around the world
- test production
- pretesting
- analysis and review
- item banking.

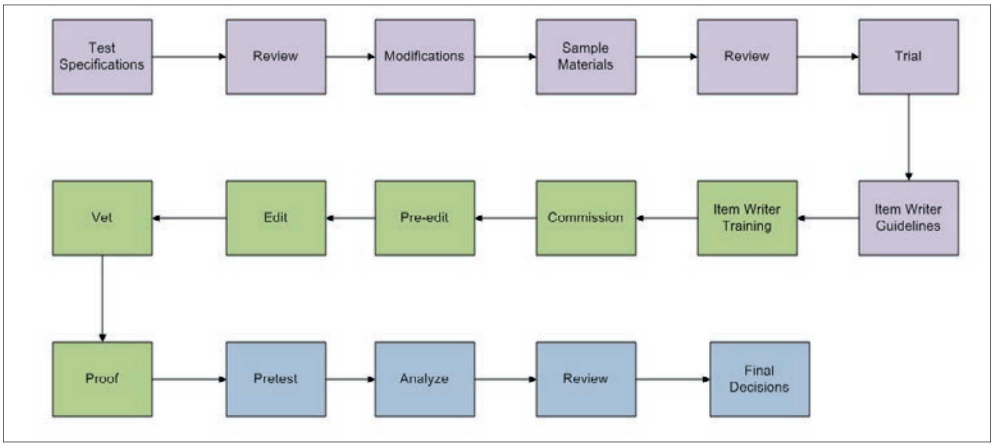


Figure 1 – The test development process

The Oxford Test of English reflects current language teaching and learning methodology. The test is designed to emulate the kinds of tasks that language learners encounter outside of test and classroom settings so that users of test results can be confident that test takers are able to perform real-world tasks.

4.1 Test design

The first phase of test development involves producing comprehensive test specifications. The specifications detail the test format, the content for each of the modules and each of the tasks contained within them. Well-crafted specifications communicate the test designers' vision, underpinning the consistency of measurement (i.e. reliability) across modules; enhancing the quality of the test across administrations and helping to ensure that decisions made based on test scores will be fair and valid.

In creating the specifications for the Oxford Test of English, OUP worked closely with institutions, teachers and learners to ensure that the test met their needs, while making certain that the test was also aligned to OUP's approach to language teaching, learning and assessment.

The specifications for the Oxford Test of English were derived from:

- level and domain descriptions in the CEFR: each task in the test is related to one or more CEFR Can Do descriptors
- communicative teaching practice
- course outlines and content from OUP teaching materials.

The test is designed to cover as wide a range of domains as possible within the confines of a two-hour administration.

Independent language-testing professionals were invited to comment on the draft specifications to help ensure appropriate coverage of domains and levels. These draft specifications were reviewed by an internal OUP panel and revised ahead of the production of sample materials. The specifications were then reviewed a second time, along with these sample materials, and further modifications were made.

Experienced item writers were commissioned to draft item writer guidelines for each module, based on the specifications and sample materials. These guidelines help our item writers to produce comparable, good-quality tasks to ensure consistency across different instances of the test and to ensure that tasks continue to reflect the intentions of the designers.

A team of item writers was trained to write an initial set of test materials. These fed into small-scale trialling in which groups of students were asked to take these tasks and provide feedback on the experience. Another round of minor revisions was then made based on the comments from other item writers and from trial students. Further sets of materials were then commissioned. These were pretested more extensively on representative samples of students in a range of countries worldwide.

4.2 Test format

All modules are delivered online so the test format was developed to reflect modern communication methods and includes task types not usually covered in traditional paper-based tests. Examples of this include an email activity in the Writing module and leaving a voicemail message in the Speaking module. Online delivery also meant that aspects of language proficiency that cannot easily be tested in paper-based tests could be incorporated, such as timed reading tasks. By allocating specific times to tasks it is possible to differentiate between speed, or expeditious, reading activities and careful reading exercises which require more time. Efforts have also been made to tap into inferred or pragmatic meanings, as well as testing more concrete understanding. A key element of the test has also been to ensure that the CEFR is covered, not just in terms of level, but also with regard to the breadth of domains covered in each skill.

The test is broken up into four modules which can be taken together in one sitting or individually. All four modules are timed, and test takers move from task to task either by selecting a 'next' button on completion of a task, or by being automatically moved to the next task at the end of the allotted time. Table 1 shows an overview of the Oxford Test of English.

Table 1: Oxford Test of English overview

Module	Part	No. tasks	No. items	Structure	Timing
Speaking	Part 1	2	6 (+ 2 unassessed)	Interview: eight spoken questions on everyday topics	Approx. 15 minutes
	Part 2	2	2	Two voicemails with spoken and written input	
	Part 3	1	1	A talk on an issue or scenario, with spoken and written input and picture prompts	
	Part 4	1	6	Six spoken questions related to the theme of the Part 3 talk	
Listening	Part 1	5	5	Five discrete short monologues/dialogues with picture options, each with one question	Approx. 30 minutes
	Part 2	1	5	A longer monologue with a note-completion task	
	Part 3	1	5	A longer dialogue with a task focusing on identifying opinions	
	Part 4	5	5	Five discrete short monologues/dialogues with text options, each with one question	
Reading	Part 1	6	6	Six short texts from a variety of sources, each with one question	35 minutes
	Part 2	1	6	Six texts, profiling people, are matched to four descriptions	
	Part 3	1	6	Six extracted sentences are inserted into a longer text	
	Part 4	1	4	A longer text with four questions	
Writing	Part 1	1	1	Email (80–130 words)	45 minutes
	Part 2	1	1	Essay (100–160 words) OR Magazine article or Review (100–160 words)	

4.2.1 Speaking module

There are four parts in the Speaking module.

In Part 1, test takers are asked to respond to eight spoken single-sentence questions on everyday topics. The first two questions are for practice purposes and are not assessed.

In Part 2, test takers are required to leave two voicemail messages.

In Part 3, test takers give a one-minute talk based on visual and audio prompts.

In Part 4, test takers answer six audio questions based on the topic of the talk presented in Part 3.

In the Speaking module, test takers wear a headset and speak into a microphone to answer questions delivered by computer. A clock displayed on the screen shows how much time is available to answer each question. Preparation time is given for the voicemails in Part 2, and the talk in Part 3.

Input is either audio-only (i.e. the text of the task is heard, but not shown on screen) or audio-written (i.e. the text of the task is heard *and* shown on screen). Where preparation time is given, this is after the task has been presented and before the test taker has to begin speaking. Table 2 shows a summary chart of the tasks in the Speaking module.

Table 2: Overview of the Speaking module

Part	No. tasks	No. items	Structure	Testing focus
Part 1	2	6 (+ 2 unassessed)	Interview Answering eight spoken single-sentence questions on everyday topics Questions 1 and 2 are always the same and given to all test takers Questions 3–5 are topic related Questions 6–8 are topic related (on a different topic to questions 3–5) **Audience:** the audience is the interviewer/assessor **Preparation time:** none **Response time:** Questions 1 and 2: 10 seconds per question Questions 3–8: 20 seconds per question	• responding to questions • giving factual information • expressing personal opinions on everyday topics
Part 2	2	2	Voicemail message Leaving two voicemail messages **Voicemail 1:** test taker leaves a voicemail Audio-visual input consisting of a situation with three prompts requiring the test taker to leave a voicemail **Audience:** the audience is specified in the task, and the relation to that audience may be informal (e.g. friend) or neutral (e.g. shop manager) **Preparation time:** 20 seconds **Response time:** 40 seconds **Voicemail 2:** test taker replies to a voicemail Audio-visual input consisting of a situation with three prompts, plus audio-only input (in the form of a voicemail which the test taker hears) requiring the test taker to leave a voicemail **Audience:** the audience is specified in the task, and the relation to that audience is informal (e.g. friend) **Preparation time:** 20 seconds **Response time:** 40 seconds	• organizing and sustaining extended discourse • sociolinguistic appropriacy • sustaining relationships
Part 3	1	1	Talk Audio-visual input in the form of a rubric and four photo prompts on an issue (e.g. what things are important for a happy life) or a scenario (e.g. how a language school can attract more students) on which the test taker gives a talk **Audience:** the audience is specified and is typically the test taker's classmates **Preparation time:** 30 seconds **Response time:** 1 minute	• organizing and sustaining extended discourse • describing • comparing and contrasting • speculating • suggesting

Part 4	1	6	Follow-up questions Answering six audio-only single sentence questions related to the Part 3 talk **Audience:** the audience is the interviewer/assessor **Preparation time:** none **Response time:** 30 seconds per question	As in Part 3, plus: • responding to questions • expressing, justifying and responding to opinions • expressing feelings

4.2.2 Listening module

There are four parts in the Listening module.

In Part 1, test takers listen to five audio recordings and, choosing from a set of options, select one picture to represent the overall meaning or specific detail of each recording.

In Part 2, test takers listen to an informational/descriptive monologue and complete a set of notes consisting of five three-option multiple-choice items.

In Part 3, test takers listen to a longer dialogue and match five statements to the speaker who expresses them.

In Part 4, test takers listen to five recordings and answer one question per recording.

The timing of all parts of the Listening module is predetermined. In each part, test takers hear each recording twice and are given a set time to check their answers before the test automatically progresses to the next recording. Table 3 shows a summary chart of the tasks in the Listening module.

Table 3: Overview of the Listening module

Part	No. tasks	No. items	Structure	Testing focus
Part 1	5	5	Multiple choice – picture options Five discrete short monologues/dialogues with picture options Five three-option multiple-choice questions Time to check answers: 10 seconds	Listening to identify: • specific information
Part 2	1	5	Note completion A longer monologue with a note-completion task Five three-option multiple-choice questions Time to check answers: 15 seconds	Listening to identify: • specific information
Part 3	1	5	Matching opinions with people who say them A longer dialogue with a task focusing on identifying opinions Five three-option multiple-choice questions Time to check answers: 15 seconds	Listening to identify: • stated opinion • implied meaning
Part 4	5	5	Multiple choice Five discrete short monologues/dialogues with text options Five three-option multiple-choice questions Time to check answers: 10 seconds	Listening to identify: • attitude/feeling/ opinion • gist • function/reason/ purpose • speaker relationship • topic • type/genre

4.2.3 Reading module

There are four parts in the Reading module.

In Part 1, test takers read six short texts from a range of genres and answer one three-option multiple-choice question on each text.

In Part 2, test takers must quickly read six profiles of people with requirements and match each to one of four topic-related factual texts.

In Part 3, test takers read a text from which six sentences have been removed, leaving gaps. Test takers choose missing sentences from a list and insert one into each gap.

In Part 4, test takers read a text and answer four three-option multiple-choice questions about the content.

All texts used in the Reading module are based on authentic material intended to be of relevance or interest to a general readership. Texts may be formal, neutral or informal in register.

The time allowed for completion of each task in the Reading module is predetermined. If the test taker does not complete the task within the allotted time, the system will automatically progress to the next task. Table 4 shows a summary chart of the tasks in the Reading module.

Table 4: Overview of the Reading module

Part	No. tasks	No. items	Structure	Testing focus
Part 1	6	6	Multiple-choice questions on short texts Six short texts from a variety of sources including: adverts, blogs, emails, notes, notices and text messages Six discrete three-option multiple-choice questions Time to process the texts and complete the tasks: 1 minute 20 seconds per task (8 minutes in total)	Reading to identify: • main message • purpose • detail
Part 2	1	6	Multiple matching Matching six profiles of people with requirements (e.g. requirements for a particular type of holiday) to four descriptions (e.g. of four different kinds of holiday) Texts from brochures, advertisements, magazine articles Six multiple-matching questions Time to process the texts and complete the task: 8 minutes	Expeditious reading to identify: • specific information • opinion and attitude
Part 3	1	6	Gapped text Six extracted sentences are inserted into a longer text Texts are from newspaper and magazine articles Six text-completion questions Time to process the text and complete the task: 11 minutes	Reading to identify: • text structure • organizational features of a text
Part 4	1	4	Multiple-choice questions on longer texts Four three-option multiple-choice questions Texts are from newspaper and magazine articles Time to process the text and complete the task: 8 minutes	Reading to identify: • attitude/opinion • purpose • reference • the meanings of words in context • global meaning

4.2.4 Writing module

There are two parts in the Writing module.

In Part 1, test takers read and respond to an input email. Responses are either informal or neutral and need to include three points from the input.

In Part 2, there is a choice of either writing an essay or a magazine article/review.

In both parts, test takers type their responses. The tasks specify a target audience and a minimum and maximum word count. There is an automatic word-count facility. Test takers will be penalized if their responses are under length.

There is a clock so that test takers always know how much time they have remaining for each part. Table 5 shows a summary chart of the tasks in the Writing module.

Table 5: Overview of the Writing module

Part	No. tasks	No. items	Structure	Testing focus
Part 1	1	1	Email 80–130 words Test taker responds to an email There are three points which the test taker must include in their email The response may be informal or neutral in tone Time to process the task and complete the response: 20 minutes	• giving information • expressing and responding to opinions and feelings • transactional functions such as inviting/requesting/suggesting
Part 2	1	1	A choice of writing tasks: an essay or a magazine article/review	
			Essay 100–160 words Writing an essay on a topic typical of classroom discussions Time to process the task and complete the response: 25 minutes	• expressing and responding to opinions • developing an argument
			or Magazine article/Review 100–160 words Writing a general article (such as the profile of a famous sports person) or writing a review (such as a review of a website) The target reader is usually an English teacher Time to process the task and complete the response: 25 minutes	• describing • narrating • expressing feelings and opinions • recommending

4.3 Test production

Before test tasks are accepted for use in the Oxford Test of English, procedures are systematically followed to ensure optimum test item quality. Rigorous adherence to such procedures helps to strengthen test quality and provides evidence that important decisions about learners' language proficiency, based on their test scores, will be valid and fair.

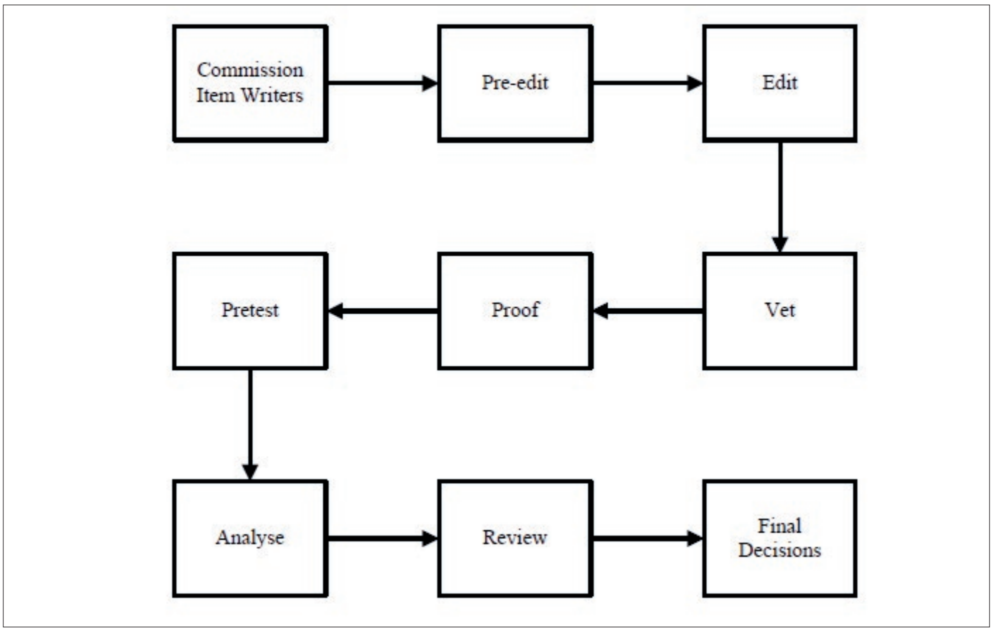


Figure 2 – The test development process

Our quality assurance process involves a number of steps. These include pre-editing, editing, vetting and proofreading before material is pretested.

Teams of item writers, led by a team leader (an expert item writer), are commissioned to work on each test module. The initial commissioning of materials is followed by a pre-editing meeting. A panel of experts reviews the materials to ensure that they closely adhere to test specifications and item-writing guidelines. The panel asks for amendments and the materials are returned to the writers to make the required changes. Once all changes have been made, the materials are further scrutinized and refined in an editing meeting.

Changes made in editing meetings are then registered on the item database, at which time the materials are vetted by an external content expert. This step provides an independent view of the material and identifies any further improvements to the task. The vetter also helps to detect: (1) whether testing points are biased towards certain language groups or cultures; (2) if items are levelled appropriately across tasks; (3) the degree to which test content is accessible on a global level; and (4) whether the test items include any unwanted taboo topics (for example, alcohol and serious illnesses). This activity safeguards against threats to test fairness.

At this point, additional materials such as audio files and graphics are added. The tasks are then proofread for instances of formatting issues and typographical errors.

The purpose of the next step in the process – pretesting – is to determine the difficulty and effectiveness of the items for use in the official, or 'live', Oxford Test of English. Students who participate in pretesting sessions are representative of the same population of students who are targeted to take the Oxford Test of English.

Data from pretesting sessions is analysed by a team of research and validation experts who employ both quantitative and qualitative methods to determine item levelling, the quality of the item options, and fit statistics for the items across tasks and levels. The statistical output, generated by the analyses, are then used for a substantive review by a panel consisting of specialists from OUP and external experts. Following pretesting and review, materials may be accepted for use in the test, sent back to item writers to be rewritten and re-pretested, or rejected and discarded.

5 Alignment to the CEFR

The CEFR is now recognized around the world as a key framework for interpreting language proficiency. Many institutions base materials, teaching programmes and tests on the CEFR levels. In developing the Oxford Test of English, every effort has been made to ensure alignment to the CEFR.

The content of the Oxford Test of English is specifically designed to elicit performances at the following levels of proficiency: CEFR levels B2, B1, and A2. This means that a test taker taking the Oxford Test of English can receive one of four results: B2, B1, A2 or Below A2. Test takers who score below level A2 receive the result 'Below A2'. This grade indicates that they are below the levels reported in the test and that we cannot ascribe a specific CEFR level to their performance. Further information on score reporting can be found in Section 8.

The CEFR has been embedded in the development of the Oxford Test of English through a range of activities. These include:

(1) employing CEFR **Can Do statements** in the test design

(2) surveying **OUP course materials** at each of the CEFR levels

(3) conducting **data analyses** on pretested items

(4) **aligning the Oxford Test of English scale** to the Oxford Online Placement Test (OOPT). Pollitt (2009) refers to work done to align OOPT to the CEFR

(5) conducting complementary **standard-setting activities** based on the Council of Europe's *Manual for Relating Language Examinations to the Common European Framework of Reference for Learning, Teaching, Assessment* (2009), henceforth referred to as 'the CEFR Manual,' to align test items to the CEFR across four modules.
A brief explanation of these activities is presented below.

(1) Can Do statements

In the design of the Oxford Test of English modules, careful attention was given to embedding links, in the form of descriptors, between the CEFR and the test items. A great effort was made to familiarize OUP item writers, item writer trainers, vetters and assessors with the CEFR with a view to linking the test specifications, item-writing guidelines and ultimately the test items to targeted CEFR Can Do statements.

(2) OUP course materials

In the development of the test specifications for the Oxford Test of English, OUP surveyed the grammatical features, degree of syntactic complexity and frequency of the lexis typically featured in Oxford University Press ELT coursebooks. On the basis of this analysis, item types and item content were identified at each CEFR level. Such findings fed into the design of the test, which benefitted from both the common understanding of levels provided by the CEFR, and from OUP's long-term practical engagement in producing English language education materials.

(3) Data analysis of pretested items

The test has been pretested around the world with over 10,000 students across thirty-seven countries from a wide number of first-language backgrounds at each of the targeted CEFR levels. Pretesting provides a good deal of information related to overall item quality (such as the quality of the item options), the performance of the test takers and the extent to which new test items could be scaled to the intended CEFR levels. Using Rasch analysis to evaluate objectively marked test tasks, a difficulty scale was plotted for the Oxford Test of English items based on Oxford Online Placement Test (OOPT) anchors [see (4) below]. Inferences about the CEFR levels can be made from such empirically-derived analyses.

(4) Aligning the Oxford Test of English to the Oxford Online Placement Test

To provide evidence of how well the Oxford Test of English is scaled to the CEFR levels, an *external-anchor design* was selected. Items from the Oxford Online Placement Test that had previously been related to the CEFR were administered to test takers as 'anchor' items alongside new material from the Oxford Test of English. An external-anchor design is often used in equating or scaling studies in which certain items link the performance of test takers across two test instruments which measure closely related knowledge and skills. (Dorans et al., 2010).

Through a series of statistical analyses, it was found that the Oxford Test of English functions on a similar scale to that of the Oxford Online Placement Test, thus providing evidence that if test takers took both tests, their test results could be interpreted on a shared scale. In other words, this provides further evidence that the Oxford Test of English and the Oxford Online Placement Test both map test takers to the CEFR in a similar way.

(5) Standard-setting activities

To strengthen inferences made from the data-driven analyses (in (3) and (4) above), a number of additional steps have been taken to ensure that the test items are appropriately aligned with the CEFR levels. Standard-setting (or benchmarking) activities were conducted to complement the pretesting-review process. Benchmarking activities, adapted from the CEFR Manual, are conducted with independent expert raters in a multi-step process:

(a) The independent experts attend a series of webinars which provide a macro- and micro-view into what the learners at each CEFR level 'can do' and what the test tasks are designed to measure.

(b) They are provided with the test specifications and item-writing guidelines.

(c) They are shown numerous examples of the test tasks from each of the modules, at varying CEFR levels, after which they are polled to determine their level of agreement. This results in the assignment of a CEFR level estimate for that item.

(d) An arbiter collects the poll results and instigates a discussion when rater disagreement requires additional adjudication. Several rounds of adjudication can occur before benchmark estimates can be established for each item.

(e) After benchmarking activities are completed, additional analyses are conducted to adjust the calibration of the benchmarked items and reconcile these with the previously pretested items. The alignment of benchmarking results with pretesting difficulties allows us to identify cut points for the CEFR levels on the Oxford Test of English scale at the B2, B1, and A2 levels.

The above procedures all contribute to the alignment of the Oxford Test of English to the CEFR, and provide evidence for the Oxford Test of English score reporting scale (see Section 8).

6 Test delivery

Unlike more traditional paper-based or linear online tests, the Oxford Test of English does not have fixed test versions in which all test takers encounter the same set of questions. Instead, it operates using an item bank and a series of selection rules. An item bank is a large collection of test questions or items that can be used during the test. The large number of items helps to ensure that different test takers using the test at the same time receive different sets of questions. The Listening and Reading modules are computer adaptive, so the tasks adapt to the ability level of the test taker. The test selection rules determine which items are presented to each test taker, for example, 'choose five Part 1 Listening items'. Each item presented to the test taker is drawn from the bank using the selection rules and an algorithm which calculates the estimated ability of the test taker and the appropriate difficulty of the next task to be presented. A randomness element is also factored into the selection of tasks, so that each test taker receives their own unique version of the test. The Speaking and Writing modules are not adaptive, but do exploit the randomness element. This approach has several advantages over traditional linear session-based tests. As test takers do not receive the same set of items, test security is improved, allowing the Oxford Test of English to be used on an on-demand basis, rather than limiting delivery to scheduled sessions. And, as the test is delivered wholly online, no materials need to be transported to test centres and stored on site, which also increases security. The item bank is refreshed on a regular basis to ensure that items do not become over-exposed. Finnerty (2015) gives further details about the advantages and workings of computer-adaptive testing (CAT).

The Oxford Test of English can only be administered by approved institutions (test centres), which are subject to ongoing quality-control checks and audits. Test centres have to provide evidence that they meet technical requirements and have the appropriate facilities and suitable staff to administer the test. Requirements for test centres are detailed in the *Oxford Test of English Test Centre Handbook*.

Once approved, a test centre can purchase test licences to run the test. The test centre then selects the date or dates on which they wish to run the test and allocates licences to that session. The Oxford Test of English can be taken on any date, though OUP usually requires fourteen days' notice of a test session – this ensures that sufficient assessors are allocated for the marking of Speaking and Writing modules. The Oxford Test of English is usually taken as a complete test (all four modules are administered in the course of a session), but test centres may choose to run sessions for single modules or any combination of modules. It is also possible for test takers to choose to resit individual modules, rather than resitting the whole test.

7 Accessibility

Oxford University Press is committed to providing accommodations to make the Oxford Test of English accessible to learners with special requirements where possible. Whilst there are some limitations to the range of accommodations that can be provided in an online test, OUP, as part of long term roadmap, will be adding additional functionality to its assessment system over the coming years to accommodate an increasing range of test taker special requirements. In the first phase of test launch, the following accommodations will be available in every test centre:

- additional time for Reading and Writing modules
- a range of colour contrast options
- increased font size.

Wherever possible, test centres will also provide the following to accommodate special requirements:

- less ambulant building access
- separate test sessions
- extended breaks between modules
- extra invigilation support.

Applications for special requirements are made by the test centre on behalf of the test taker. The option to adjust colour contrast and font size are applied to the test taker's test profile by OUP and *do not require* supporting medical documentation. Requests for additional time, extended breaks or a separate test session *need to be accompanied* by the appropriate medical certificate.

8 Test marking and scoring

The Reading and Listening modules employ an adaptive algorithm. Depending on whether a correct or incorrect response is received for each task, the system increases or decreases the difficulty of the items as the test progresses. Responses for Reading and Listening are marked by computer and the ability of the test taker is estimated according to the responses given in relation to the difficulty of the questions presented. The Oxford Test of English employs the *Weighted Maximum Likelihood Estimation* (Warm, 1989) in its test algorithm. The equation in this formula uses the test taker's responses to items of different Rasch difficulties to estimate their ability at each decision point, i.e. at the end of each item or set of items.

As the test progresses, the estimate of the test taker's ability is refined using additional information from each item or set of items and the statistical error associated with the estimate is reduced. To ensure that each test taker has the same test experience, the Oxford Test of English delivers a standard test format to each test taker. That is, all test takers receive the same task types and the same number of items. The final ability estimate is derived once the complete set of test items in a module has been delivered. Ability estimates are then converted to a standardized score and this is also reported in terms of a CEFR level.

Speaking and Writing tasks are selected at random from the item bank, according to a pre-defined number and order of tasks, and the responses are returned online and sent electronically to trained assessors who mark them according to analytical criteria derived from the CEFR level descriptors. All Speaking and Writing assessors have significant English language-teaching experience and recognized English language-teaching qualifications.

Assessors follow a standardized training and certification process before being allowed to participate in marking. Their marking is then monitored to ensure consistency over time.

Automated quality assurance monitoring is carried out using Speaking and Writing responses which have been marked previously by a number of experienced raters (and so have agreed benchmark ratings). These 'seeded' responses are included together with the new test responses to check that the assessors continue to be accurate in their marking to within set tolerances. All responses are anonymous (so assessors are unaware whether the responses they are marking have already been rated). Assessors whose marking falls outside of agreed tolerances are removed from the marking process and asked to complete a re-standardization process, after which they can resume marking. Assessors who do not successfully complete re-standardization are permanently withdrawn from marking.

9 Results reporting

Performance on the Oxford Test of English is reported in terms of standardized scores on a scale ranging from 0 to 140. Standardized scores are independent of test sessions and give a standard reference point for students taking the test on different occasions.

Results are also displayed as a bar chart, showing how performance on the test relates to the relevant CEFR levels.

Table 6 shows the relationship between the Oxford Test of English scale and the CEFR.

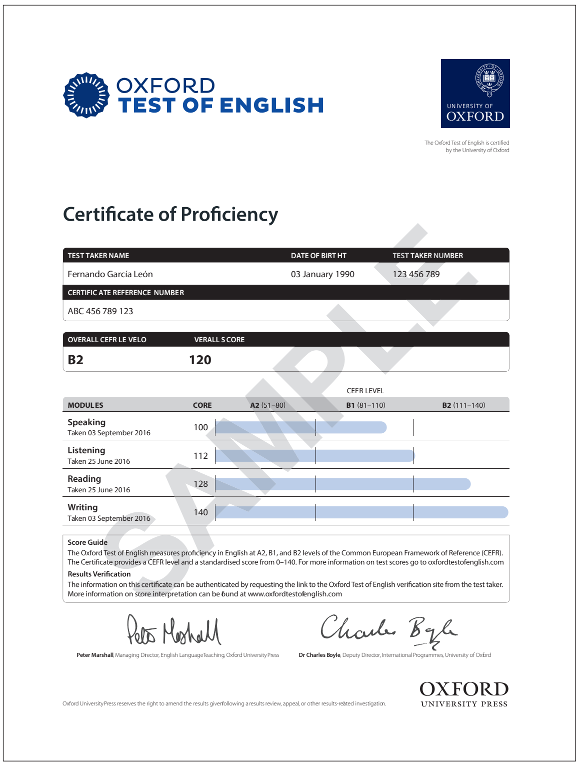
Table 6: The Oxford Test of English scale and the CEFR

CEFR band	Oxford Test of English score range
B2	111–140
B1	81–110
A2	51–80
Below A2	0–50

The Oxford Test of English reports scores between CEFR levels B2 and A2. This means that although a test taker's responses may indicate performance that is above B2 level, a test taker cannot receive a test score above B2. The rationale for this is that the test taker has received tasks designed for CEFR levels B2, B1, and A2, so we cannot be certain how they would have performed on tasks designed for C1 or C2 test takers. The Oxford Test of English does, however, give an indication of 'Below A2' performance. Below A2-level performance means that a test taker is not at the level the test was designed to measure and that no precise statement of level can be made. For the objectively marked Reading and Listening modules, the final ability estimates obtained through the test algorithm are converted to standardized scores and these are used in determining the CEFR levels. For Speaking and Writing, marks are awarded by assessors, using the analytical marking criteria. These marks are then converted into standardized scores.

Test takers receive a standardized score and CEFR level on a Module Report Card for each module taken. If a test taker completes all four modules, they also receive an overall score and CEFR level on an Oxford Test of English Certificate. The overall score is calculated as an average of the scores obtained in each of the four modules.
Figure 3 shows a sample test certificate.

Figure 3: Sample test certificate



OXFORD
TEST OF ENGLISH

UNIVERSITY OF
OXFORD

The Oxford Test of English is certified
by the University of Oxford

Certificate of Proficiency

TEST TAKER NAME	DATE OF BIRT HT	TEST TAKER NUMBER
Fernando García León	03 January 1990	123 456 789
CERTIFIC ATE REFERENCE NUMBER		
ABC 456 789 123		

OVERALL CEFR LE VELO	VERALL S CORE
B2	120

CEFR LEVEL

MODULES	CORE	A2 (51–80)	B1 (81–110)	B2 (111–140)
Speaking Taken 03 September 2016	100			
Listening Taken 25 June 2016	112			
Reading Taken 25 June 2016	128			
Writing Taken 03 September 2016	140			

Score Guide
The Oxford Test of English measures proficiency in English at A2, B1, and B2 levels of the Common European Framework of Reference (CEFR).
The Certificate provides a CEFR level and a standardised score from 0–140. For more information on test scores go to oxfordtestofenglish.com

Results Verification
The information on this certificate can be authenticated by requesting the link to the Oxford Test of English verification site from the test taker.
More information on score interpretation can be 6und at www.oxfordtestofenglish.com

Peter Marshall

Peter Marshall, Managing Director, English Language Teaching, Oxford University Press

Charles Boyle

Dr Charles Boyle, Deputy Director, International Programmes, University of Oxford

Oxford University Press reserves the right to amend the results givenfollowing a results review, appeal, or other results-related investigation.

OXFORD
UNIVERSITY PRESS

10 Results reviews and appeals

However effective a testing programme may be, test takers or other stakeholders may wish to challenge or appeal their result and transparent procedures must be open to them. There is a two-stage process for challenging a result on the Oxford Test of English: results review and appeal.

For a results review, the test results for one or more modules are checked or re-marked. For Speaking and Writing, a results review involves a re-mark of the responses. This is done by inviting senior assessors to re-mark the module in question. If the re-mark results in a score that improves the module or overall CEFR level, the results enquiry is upheld and the test taker receives a replacement result.

For Reading and Listening, the results review will involve a results check. As Reading and Listening are both marked by computer, there is no scope for re-marking as the re-mark result would be identical to the original result. However, a check is made by OUP on the tasks presented to the test taker to ensure that they received tasks at the appropriate level and that their ability estimate was correctly calculated. If an error is identified with the result, a decision will be made as to whether a revised result can be issued or whether the test taker should be given the opportunity to resit the module.

A test taker can also request an appeal via their test centre. An appeal differs from a results review in that an appeals panel, which is entirely independent of OUP, undertakes the investigation of the test taker's responses and marks to ensure that all appropriate steps have been taken in reviewing the result. The Oxford University Department for Continuing Education (OUDCE) acts as the independent appeals body for the Oxford Test of English. An appeal must be preceded by a results review.

An administrative fee is charged for all results reviews and appeals, but the fee is refunded if the review results in a change of CEFR level for either a module or the whole test, or if the appeal is upheld. All results reviews and appeals are processed on behalf of the test taker by the test centre at which the test was administered.

11 Test monitoring, impact and review

The development and administration steps outlined above have been designed to ensure that every administration of the Oxford Test of English provides reliable results that serve as a valid basis for decision-making.

To ensure that the Oxford Test of English continues to fulfil its stated purpose, and to seek opportunities for further improvements in quality, OUP monitors test administrations and carries out analyses of the performance of test materials, test takers and assessors at regular intervals.

Data from test administrations and feedback from assessors and stakeholders will lead to opportunities to review the test and improve its format and content in the light of experience over future years.

12 Acknowledgements

Oxford University Press would like to thank the following consultants for their input into and/or reviews of the design and development of the Oxford Test of English:

Professor Charles Alderson

Dr Nathan T. Carr

Dr John Field

Professor Anthony Green

Professor Claudia Harsch

Dr Alastair Pollitt

Dr Philida Schellekens

Dr Norman Verhelst

13 References

Bachman, L. F., and Palmer, A. S. (2010). *Language Assessment in Practice.* Oxford: Oxford University Press.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Council of Europe (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): A Manual.* Strasbourg, France: Language Policy Division.

Dorans, N., Pommerich, M., and Holland, P. (eds.) (2010). *Linking and Aligning Scores and Scales (Statistics for Social and Behavioral Sciences).* New York: Springer Publishing Company.

Finnerty, C. (2015). 'The CAT is out of the bag'. *Modern English Teacher*, 24 (3): 15–17.

Pollitt, A. (2009). *The Meaning of OOPT Scores.* https://www.oxfordenglishtesting.com Oxford: Oxford University Press.

Warm, T.A. (1989). 'Weighted Likelihood Estimation of Ability in Item Response Theory'. *Psychometrika*, 54(3): 427–450.