Pearson English

Benchmark Test

Test Specification and Validation Report

3 November 2020 | v 2.0

Benchmark

Introduction

Benchmark Test: An overview

The Pearson English Benchmark Test is a course agnostic online assessment of English language proficiency for individuals and groups of students. It can be used to as a one off test to assess Reading, Writing, Speaking and Listening skills or to measure progress over time. The test measures a student's English language competency and reports a CEFR score from pre-A1 to C2, together with a GSE score between 10 and 90.

The score report provides the overall test result, with skill scores, performance summaries, and recommendations for future study. Results can be reported at both the group and individual level. The assessment is delivered digitally through the Pearson English Test Hub, which also stores and displays the results of the test.

The purpose of the test

The test is normally used in a learning context in conjunction with relevant materials courseware and other formative assessment tasks. It provides detailed information to a teacher about individual students and groups of learners who are studying English. The information provided includes:

- An overall Global Scale of English (GSE) score and CEFR band for each student
- A profile of sub-skill scores for a group showing areas of strengths or weaknesses for the group
- A profile of sub-skill scores for each student which show the strengths and weaknesses for each student
- Comparison of scores from one test administration to another so progress, or lack of it, can be viewed. This is available at the group and individual level.
- Performance summaries, recommendations for further study, and links to Pearson courseware and GSE Learning Objectives

This information allows the teacher to measure student progression and make decisions about adapting learning material to suit the level of learners, as well as providing extension activities where needed. It also allows the teacher to tailor the learning program to particular learners, giving extra support and input where required.

Who is it for?

The test is designed for secondary and adult learners who are aged 14 or older. Benchmark Test can be used alongside any adult or upper secondary course and is intended to be used with comprehensive integrated skills courses not short or partial courses.

Why take an integrated skills test?

Some of the questions Benchmark Test uses test a single skill such as speaking or writing. When assessing these skills, we test traits such as pronunciation and fluency, the ability to argue as well as written conventions along with grammar and vocabulary.

A number of the questions on the test are integrated skills questions. These questions test more than one skill at the same time. Using integrated skills questions means that Benchmark Test is a better test of a learner's English. In real life and in the classroom learners use more than one skill to complete communicative tasks. To order something in a restaurant we need to listen and speak, to take notes in a classroom we need to listen and write. Integrated skills questions test

how well learners can use the skills they have learnt and practised in the classroom and used in real life.

Test Design

Benchmark Test is designed specifically to measure progress in language proficiency. The test construct is based on actionable learner outcomes as embodied in can-do statements in the Common European Framework of Languages and the Global Scale of English General learning objectives. It is built on the body of applied linguistic research of the last 50 years which prioritises the ability to use language in context rather than just knowledge of the language. In order to use language effectively it is assumed that learners require certain knowledge of the systems of language such as grammar, vocabulary and phonemic systems.

The test explicitly measures language as a unitary trait. It also provides a breakdown of sub-scores for the convenience of users to provide some insights into the relative strengths and weaknesses of students.

The test suite contains 4 tests: Test A, Test B1, Test B2 and Test C. Test A assesses at CEFR A1 and A2, and Test C at C1 and C2. The tests use fixed, linear forms. The total time of the test is approximately 45 minutes, although this may vary slightly according to level. Each test has three parts:

Part 1 - Part 1 tests Grammar, Vocabulary and Reading. There are 8 item types in this section.

Part 2 - Part 2 assesses Speaking and Listening. There are 7 item types in this section.

Part 3 - Part 3 assesses Writing. There are two item types in this section.

Test items are primarily integrated skills and scored on the GSE scale. Speaking and Writing items utilise automated scoring technologies, described below.

Global Scale of English and the Common European Framework levels

In the following tables we define how the Global Scale of English is related to the CEFR levels. To give an impression of what the levels mean, i.e., what learners at particular levels can do, we use the summary descriptors published in the CEFR (Council of Europe, 2001, p. 24) where provided.

GSE 10–21 **Global assessment** The range on the Global Scale of English from 22 to 29 corresponds to the pre-A1 level of the CEFR. There are no Global descriptors for pre-A1, but abilities at this level can be summarised as follows:	This level of proficiency is likened to a tourist who may know some individual words but does not have enough control of language to produce full sentences and mostly communicates with words or very basic phrases. The words they do know may carry a lot of communicative meaning or be effective when used with hand gestures or when the context is very clear (e.g. pointing to an object in a shop).
GSE 22–29 **Global assessment** The range on the Global Scale of English from 22 to 29 corresponds to the A1 level of the CEFR. The capabilities of learners at Level A1 have been summarised in the CEFR (Council of Europe, 2001, Table 1, p. 24) as follows:	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

GSE 30-35 and 36-42 **Global assessment** The interval on the Global Scale of English from 30 to 35 corresponds to the lower part of the A2 level of the CEFR, while the interval from 36 to 42 corresponds to the upper part of the A2 level, which is also sometimes referred to as the A2+ level. The capabilities of learners at Level A2 have been summarised in the CEFR (Council of Europe, 2001, Table 1, p. 24) as follows:	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
GSE 43–50 and 51-58 **Global assessment** The interval on the Global Scale of English from 43 to 50 corresponds to the lower part of the B1 level of the CEFR, while the interval from 51 to 58 corresponds to the upper part of the B1 level, which is also sometimes referred to as the B1+ level. The capabilities of learners at Level B1 have been summarised in the CEFR (Council of Europe, 2001, Table 1, p. 24) as follows:	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
GSE 59-66 and 67-75 **Global assessment** The interval on the Global Scale of English from 59 to 66 corresponds to the lower part of the B2 level of the CEFR, while the interval from 67 to 75 corresponds to the upper part of the B2 level, which is also sometimes referred to as the B2+ level. The capabilities of learners at Level B2 have been summarised in the CEFR (Council of Europe, 2001, Table 1, p. 24) as follows:	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and Independent disadvantages of various options.
GSE 76–84 **Global assessment** The interval on the Global Scale of English from 76 to 84 corresponds to the C1 level of the CEFR. The capabilities of learners at Level C1 have been summarised in the CEFR (Council of Europe, 2001, Table 1, p. 24) as follows:	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
GSE 85 to 90 **Global assessment** The interval on the global scale corresponds to the level C2 on the CEFR. This is a very high level of attainment which has been summarised in the CEFR (Council of Europe, 2001, Table 1, p. 24) as follows:	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.

Test Development

The questions in Benchmark Test have been developed by international teams of writers who are very experienced in writing assessment questions. Teams are based in the UK, Australia, the USA and Hong Kong. All questions have been tagged with a Global Scale of English (GSE) level and linked to a 'can do' statement.

Once written, all questions are reviewed by the teams in the different countries. Comments and suggestions for improvement are stored with the test questions on a secure database. The questions then go through a further review by an expert panel and decisions are made on the quality of the questions; which to keep and which to reject. All questions are then thoroughly checked by Pearson staff and images and high quality recordings are added to complete the questions before they go forward to be calibrated in a large scale field test.

After the field testing, further checks are made on item quality based on the measurement characteristics of the questions. Questions are eliminated from the item pool if they are too easy or too difficult, if weaker learners get them right but stronger learners get them wrong, or if they show any bias. These checks then result in a bank of the best quality questions. Questions are selected from this bank to go into the final tests.

Field Testing

As part of the test development process, a large field test, conducted in two phases, was carried out to ascertain the appropriateness of the pool of items and to serve as a source for constructing individual test forms which would allow reliable predictions of students' ability in English. A portion of the data collected was transcribed and rated which was used to train automated scoring systems.

Field test forms were created using a linking approach. That is, the forms were linked together with sets of items that appeared on all forms. Also, during the second phase of data collection, since most candidates took two tests, the field test forms were also linked through candidates.

Learners and L1 English speakers were recruited to participate in the field test. A total of 13,073 tests were submitted during the two field test phases. The demographic for Benchmark is upper secondary and young adult. The majority of participants were aged 16 to 35. Participants were from 96 countries. The countries with the largest number of participants included; Saudi Arabia, Poland, Panama, Ecuador, The Netherlands, Argentina, Brazil, Spain, Guatemala, Japan and Thailand. As an incentive to participate, students received a year's free access to the Longman Dictionary of Contemporary English Online (LDOCE). L1 English speakers were offered an Amazon voucher.

Validity Evidence

Test Reliability

Reliability is one aspect of validity - if a candidate took a test on multiple occasions, would that person get a similar score each time? During field testing, a large number of candidates took two tests in a short period of time. The two tests were made up of different items. Presumably, little or no learning occurred between these test administrations, so the correlation of the scores from these two tests should provide a good estimate of test reliability, known as test-retest reliability. The higher the observed correlation between the two test administrations, the more reliable the test scores are. In the observed field test data, after removing test data from candidates who either did not answer a sufficient number of items, or who got extreme scores outside of the normal GSE range, the test-retest correlation was .861 (n=2,141). This observed correlation demonstrates a high level of consistency of measurement of Benchmark test administrations.

The psychometric analysis tool, called Winsteps, also yielded another measure of test reliability estimate as part of item calibration. The reliability estimate is 0.90 (n=11,908). From these two estimates, it is clear that test reliability is high.

Automated scoring validation process

From the field test data, 300 candidates were randomly selected as the validation data set. A validation data set is a group of candidates whose data are segregated out prior to psychometric analysis in order to independently test how well automated scoring models work, once they are complete. Additionally, these candidates' data were not included in the psychometric item calibration, or in the scaling onto the GSE. If the test scores for these candidates as calculated by both automated and human scoring models are highly correlated, this provides evidence that the automated scoring models will work as expected for other new candidates in the operational setting.

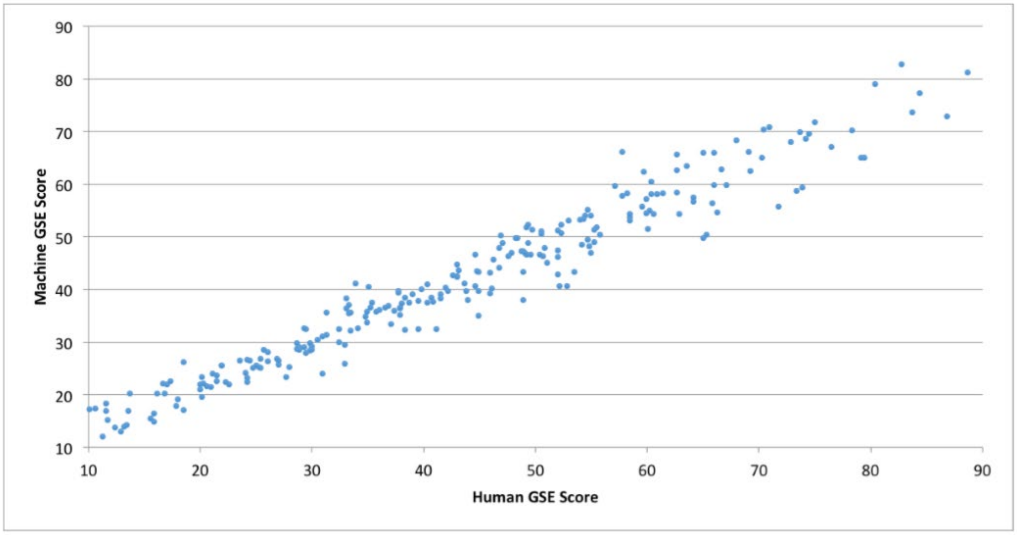
Once the automated scoring system was developed, the responses from the validation set were run through the same psychometric model to produce an Overall and six skill scores for each candidate. Those human and machine scores were then correlated to compare how similar those two kinds of scores are for each person. When candidates were identified as having extreme scores (i.e., well outside the reported score range of the GSE and not well estimated), or when they had fewer than five responses which were able to be scored in a skill area, their scores were excluded from the analyses. This reduced the n-count for the Overall score correlation to 288 candidates. The relationship between machine and human Overall scores was found to be a very strong one with a correlation of .97 (see Table 1).

Table 1. Correlations between scores using machine and human scoring methods for Overall and skill reporting areas.

Score Type	Correlation
Overall	.97
Listening	.93
Speaking	.83
Reading	.90
Writing	.99
Grammar	.97
Vocabulary	.93

Machine scoring produces scores that are nearly identical to those that a careful human rating process for many item types would require (see Figure 1).

Figure 1. Scatter plot of GSE scaled scores for validation set candidates using human and machine scoring methods.



Test reporting

A key feature of the Benchmark Test is the score report, specifically the performance summaries, recommendations, GSE Learning objectives, and relevant courseware activities. These provide more detailed feedback for teachers and students and help advise future learning.

Using a Benchmark Test assessment map, based on Pearson assessment frameworks, item types and content, the score report provides test-specific performance summaries for each skill at different CEFR levels. The performance summaries link to student and teacher recommendations as well as GSE Learning Objectives, which state what students need to do to improve i.e. reach the next CEFR band. Recommendations are course agnostic, but a range of Pearson courseware titles are available to provide links to specific activities.

The performance summaries, recommendations and recommended activities form part of the test reporting, together with overall and skill scores. Enabling skills (Vocabulary, Grammar) are reported on in the performance summaries and recommendations where relevant, but are not be given a specific GSE score. Reporting is provided at the individual and group level.

The test and reporting are course agnostic, but specific courseware activities related to the recommendations are provided for selected ELT courseware

Test Questions

What kinds of questions are in the test and what do they measure?

The test has a number of different question types. This gives learners a chance to demonstrate their English skills in different ways. There are questions where learners choose the correct option or where they write the answer into an open question. There are questions where the learner repeats or copies what has been said as well as questions where learners describe something or write a short essay. The questions are similar to the questions and tasks learners will have done in the classroom as part of their learning and so should be familiar. Not all question types appear at all levels.

Vocabulary questions

There are three vocabulary question types. Vocabulary is also tested as part of *Describe Image*, *Short Essay* and *Recall a Passage* which are integrated skills questions.

Item type	What do the learners have to do?	What is being tested?
Fill in the Table	This question asks the learner to complete a set of vocabulary items with appropriate words. The words are presented as a table of related words.	This question tests the vocabulary knowledge of the learner. It tests the words the learner knows and the accuracy of the form of the word. It tests the learner's knowledge of word families and related sets of words that they may have met in the classroom or when learning English.
Choose the Right Word or Phrase	This question asks the learner to choose the correct word to complete a number of sentences. The sentences are related by a similar theme.	This question tests the vocabulary knowledge of the learner in a written context. It tests the vocabulary the learner knows and whether they can understand the use of the vocabulary in the context of a sentence. It tests the range of vocabulary the learner knows.
Complete the Dialogue	This question asks the learner to select words from a word bank to complete a dialogue.	This question tests the vocabulary of the learner in a spoken context. It tests the vocabulary the learner knows and whether they can understand the use of the vocabulary in the context of a conversation. It tests the range of vocabulary the learner knows.

Grammar questions

There are two grammar question types. Grammar is also tested as part of *Short Essay* and *Recall a passage*.

Item type	What do the learners have to do?	What is being tested?
Choose the Right Word or Phrase	This question asks the learner to choose the correct word or phrase to complete a number of sentences. The sentences are related by a similar theme.	This question tests the knowledge of grammar of the learner. It tests the range of grammatical knowledge as well as the accuracy of grammar in a written context.
Correct the Mistake	This question asks the learner to choose the correct word or phrase to replace a mistake in a sentence	This question tests the knowledge of grammar of the learner in a written context and whether they can choose the right grammatical form in a sentence.

Reading questions

There are three reading question types. Reading is also tested as part of *Read aloud*, *Recall a passage*, and *Listen and Read* which are all Integrated Skills questions.

Item type	What do the learners have to do?	What is being tested?
Choose the Right Picture	This question asks learners to read a short text and select the best picture to match with the text.	This question tests the global understanding of short messages, notes and short pieces of writing.
Short Answer Questions	This question asks the learner to read a longer text and answer questions on the text.	This question tests the reading comprehension of the learner. It tests specific information included in the text.
Choose the Right Word or Phrase (gapfill)	This question asks learners to read a short text and select the best word or phrase to complete the text.	This question tests the global understanding of short messages, notes and short pieces of writing.

Listening questions

Listening is tested as part of *Listen and Repeat, Story Retell, Listen to the conversation, Listen and Write (Dictation)*, and *Listen and Read (Hotspots)* which are all Integrated Skills questions.

Speaking questions

There is one speaking question type which tests speaking. Speaking is also tested as part of *Read aloud, Listen and Repeat, Story retell, Listen to the conversation and Passage comprehension* which are Integrated Skills questions.

Item type	What do the learners have to do?	What is being tested?
Describe Image	This question asks the learner to look at a photograph or picture and describe what they see.	This question tests the learner's ability to speak in an extended way linking concepts and ideas. It tests the accuracy of speech including accurate grammar, pronunciation and stress as well as the fluency of the speech. It tests the use of appropriate words to describe the photograph or picture.

Writing questions

There is one question type which tests only writing. Writing is also tested as part of *Listen and Write* and *Recall a passage* which are Integrated Skills questions.

Item type	What do the learners have to do?	What is being tested?
Short Essay	This question asks the learner to write a short essay in response to a prompt. For lower levels, test takers need to write a short description of an image	This question tests global writing skills. It tests paragraph and sentence structure, the range and accuracy of the language used, the ability to structure an argument or discussion in a written context. It tests grammar and vocabulary as an essential part of writing.

Integrated skills questions

There are seven question types which measure more than one skill at the same time. These are called Integrated Skills Questions.

Item type	What do the learners have to do?	What is being tested?
Read Aloud	This question asks the learner to read aloud a sentence or short text.	This question tests accurate pronunciation and how fluent the learner is at speaking. It tests if the words in the text are understood and repeated accurately.
Listen and then Write (Dictation)	This question asks the learner to listen to a sentence or short text and write what they have heard.	This question tests listening comprehension at the word and sentence level. It tests the ability to write accurately and understand sentence structure, word order and connectors.
Listen and Repeat	This question asks the learner to listen to a sentence or short text and then repeat it	This question tests listening comprehension at the word and sentence level. It tests pronunciation and fluency. It tests if the words heard are understood and repeated accurately
Read and then Write	This question asks the learners to read a short story or short piece of factual text. The text then disappears and the learner has to reconstruct the text.	This question tests reading comprehension. It tests the ability to write accurately and understand sentence structure, word order and connectors.
Listen and Read (Hotspots)	This question asks the learner to read a text and at the same time listen to the text. The learner has to find the differences between the written text and the spoken text.	This question tests reading and listening comprehension. It tests the ability to recognise individual words in a text.
Story Retell	This questions asks the learner to listen to a short narrative and then retell the narrative using their own words.	This questions tests listening and speaking. It assesses understanding of a short narrative
Listen to the Conversation	This question asks the learner to listen to a short conversation and then answer a question about the conversation.	This question tests listening comprehension. It tests the accuracy of the listening comprehension of the learner

Question type and level

Most questions are used across all 4 levels of the test. However, some questions are more appropriate for students at lower or higher levels of proficiency. The table below shows question types in relation to four different test levels.

	Item Type	Skills covered	Test Level			
			A	B1	B2	C
Part 1	Fill in the Table	Vocabulary	■			
	Choose the Right Word or Phrase	Vocabulary	■	■	■	■
	Choose the Right Word or Phrase	Grammar	■	■	■	■
	Complete the Dialogue	Vocabulary		■	■	■
	Correct the Mistake	Grammar	■	■	■	■
	Choose the Right Picture	Reading	■			
	Short Answer Questions	Reading	■	■	■	■
	Choose the Right Word or Phrase	Reading	■	■		
Part 2	Read Aloud	Speaking & Reading	■	■	■	■
	Listen and Repeat	Listening & Speaking	■	■	■	■
	Describe a Picture	Speaking	■	■	■	■
	Story Retell	Listening & Speaking		■	■	■
	Listen to the Conversation	Listening & Speaking		■	■	■
	Listen and Read (Hotspots)	Listening & Reading	■	■	■	■
	Listen and Write (Dictation)	Listening & Writing	■	■	■	■
Part 3	Recall a Passage	Reading & Writing	■	■	■	■
	Short Writing Task (Essay or describe an image)	Writing	■	■	■	■

Sample test

Learners can take the unscored sample test to familiarise themselves with the question types in the test. There is a sample test for every Benchmark level as question types and the level of difficulty vary. The sample test is approximately half the length of the full test and contains examples of all the item types students will attempt in the scored version of the test. Teachers should run through the sample test the first-time students take a test at a particular level.