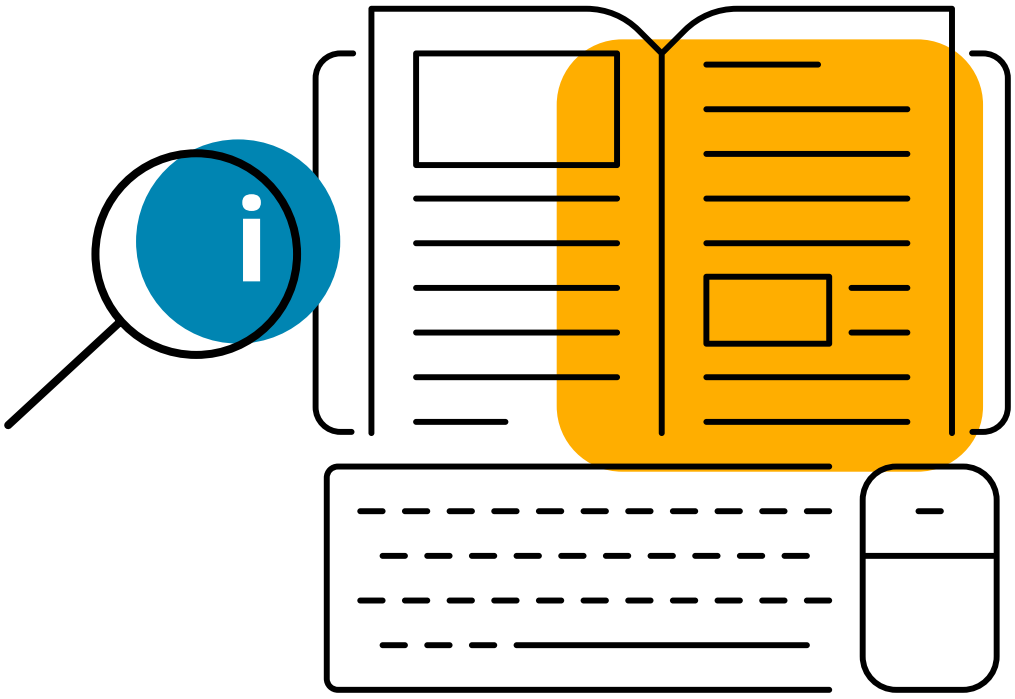
Pearson
PTE Academic



PTE Academic

Institution

Score Guide

VERSION 13 – NOVEMBER 2024

Table of Contents

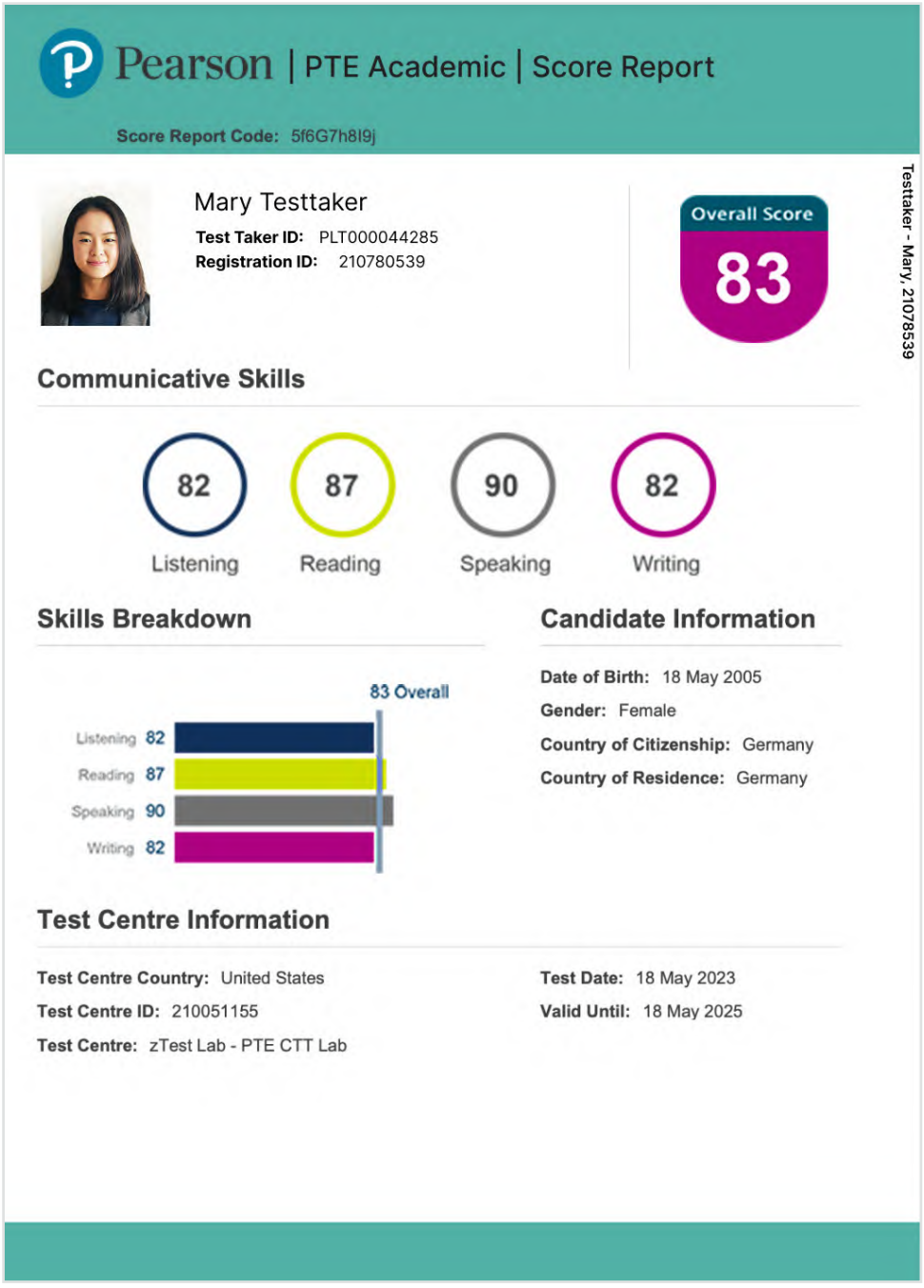
Introduction

Pearson Test of English Academic (PTE Academic) is an international computer-based English language test. It provides a measure of a test taker's language ability in order to assist education institutions and professional and government organizations that require a standard of academic English language proficiency for admission purposes.

The contents of this Guide, along with those published on **our website**, provide the only official information about PTE Academic.

1. Reported Scores

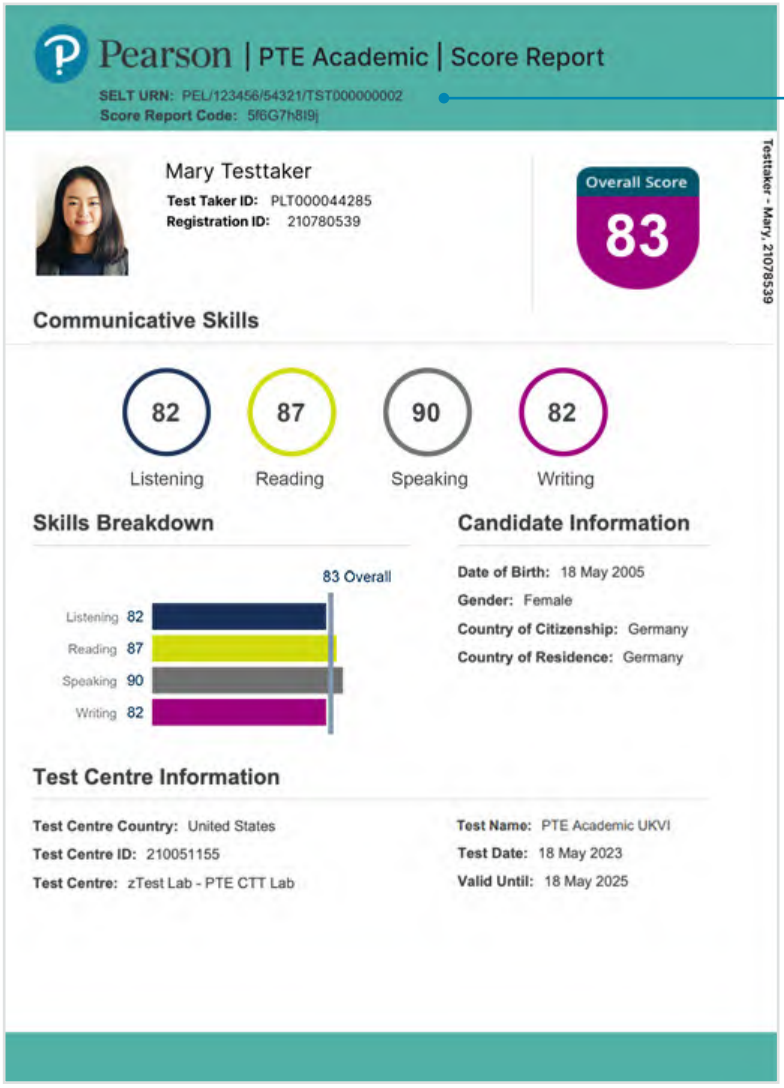
The PTE Academic Score Report consists of an **overall score** and four **communicative skills scores (listening, reading, speaking, writing)** as shown below.



Alternative versions of PTE Academic

PTE Academic UKVI

The PTE Academic UKVI test is taken for UK visas and immigration purposes. The Score Report is identical in content but contains a SELT URN number which allows the UK Government to verify the score.



SELT URN number

Overall score

The **overall score** is based on your performance across the entire test. You will complete between 52 and 64 tasks in any given test and a range of 20 different task types.

The overall score ranges between 10–90 points.

Note: *the overall score is not an average calculation of the communicative skills scores.*

Communicative skills scores

The communicative skills are **listening, reading, speaking,** and **writing** and their score range is 10–90 points.

Some tasks assess more than one skill at the same time and they are called integrated skills tasks (assessing reading and speaking, listening and speaking, reading and writing, listening and writing or listening and reading). The scores on these tasks contribute to the score of both communicative skills that are assessed at the same time.

Note: *Score Report for tests taken before November 16, 2021 included additional skills called enabling skills. These skills have been removed from the Score Report and we introduced a personalized **'Skills Profile'** to provide guidance on how to improve test takers' English proficiency.* For more information go **here**.

Enabling skills are not recommended for visa requirements or for admission purposes. Institutions should use overall score and communicative skills scores to assess test takers English proficiency levels.

2. Using PTE Academic Scores

How to use PTE Academic scores

Our experience suggests that most universities require:

Degree/Course type	Recommended cut scores
Foundation courses	minimum score of **between 36–50**
Undergraduate degrees	minimum score of **between 51–60**
Postgraduate degrees	minimum score of **between 57–67**

Each Higher Education Institution determines their own admissions criteria and as such sets their own minimum threshold for the score required to study at degree level or above. This can be based on:

1. A student's overall score
2. A student's overall score in conjunction with their communicative skills scores.

For example, institutions may:

- Set the admission requirement based on the minimum overall score alone.
- Set the admission requirement based on the minimum overall score in combination with a higher minimum on one of the communicative skills scores, because it is considered particularly important for the program the test taker wants to enter.
- Set the admission requirement based on the minimum overall score in combination with a lower minimum on one of the communicative skills scores, because it is considered less important for the program the test taker wants to enter.

Other combinations of the overall score and one or more of the communicative skills scores may be considered.

3. Concordance between PTE Academic, CEFR, TOEFL and IELTS

Based on research and empirical concordance studies, we have produced concordance tables showing the relationship between the PTE Academic test, the IELTs academic test, TOEFL iBT and the Common European Framework of Reference for Languages (CEFR). The table on the next page shows our current best estimate of concordance between PTE Academic scores and the Common European Framework of Reference for Languages (CEFR). In addition, shaded score ranges indicate the PTE Academic scores that predict some degree of performance at the next CEFR level.

Please note that any attempt to predict a score on a particular test, based on the score observed on another test, will contain measurement error. This is caused by the inherent error in each of the tests in the comparison and in the estimate of the concordance. Furthermore, tests in the comparison do not measure the same construct.

Alignment with the CEFR

To ensure comparability and interpretability of test scores, PTE Academic has been aligned to the CEFR, which is recognized as a standard across Europe and in many countries outside of Europe. In the USA, the National Council of State Supervisors for Languages (NCSSFL) has introduced the use of the LinguaFolio Self-Assessment Grid (NCSSFL, 2008), which relates language levels to the scales of both the ACTFL (American Council on the Teaching of Foreign Languages) and the CEFR.

The CEFR includes a set of consecutive language levels defined by descriptors of language competencies. The six-level framework was developed by the Council of Europe (2001) to enable language learners, teachers, universities or potential employers to compare and relate language qualifications by level.

Alignment of PTE Academic to the CEFR levels provides a means to interpret PTE Academic scores in terms of the level descriptors of the CEFR. As these descriptors focus on what an English language learner can do, scores that are properly aligned to the CEFR give educators and institutions more relevant information about a test taker's ability.

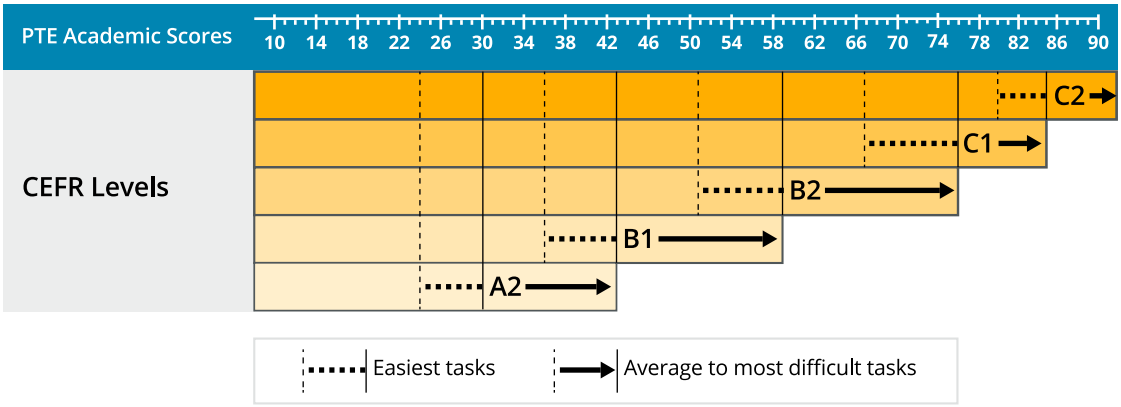
Please refer to the **Appendix** for a detailed table showing PTE Academic scores aligned to CEFR ranges.

The PTE Academic Score Scale and the CEFR

The explanation of the alignment of PTE Academic to the CEFR is that to stand a reasonable chance at successfully performing any of the tasks defined at a particular CEFR level, learners must be able to demonstrate that they can do the average tasks at that level.

As students grow in ability, for example within the B1 level, they will become successful at doing even the most difficult tasks at that level and will also find they can cope with the easiest tasks at the next level. In other words, they are entering into the B2 level.

The following diagram shows PTE Academic scores aligned to the CEFR levels A2 to C2. The dotted lines on the scale show the PTE Academic score ranges that predict that test takers are likely to perform successfully on the easiest tasks at the next higher level. For example, if a candidate scores 51 on PTE Academic, this means that they are likely to be able to cope with the more difficult tasks within the CEFR B1 level. At the same time, according to their PTE Academic score, it predicts that they are likely to perform successfully on the easiest tasks at B2.



Alignment of PTE Academic scores to the CEFR.

Below are descriptions of those threshold levels defined by success on the easier tasks for B2 and B1.

B2 Threshold Level

Has sufficient command of the language to deal with most familiar situations but will often require repetition and make many mistakes. Can deal with standard spoken language but will have problems in noisy circumstances. Can exchange factual information on familiar routine and non-routine matters within his/her field with some confidence. Can pass on a detailed piece of information reliably. Can understand the information content of the majority of recorded or broadcast material on topics of personal Interest delivered in clear standard speech.

B1 Threshold Level

Has limited command of language, but it is sufficient in most familiar situations provided language is simple and clear. May be able to deal with less routine situations on public transport e.g., asking another passenger where to get off for an unfamiliar destination. Can re-tell short written passages in a simple fashion using the wording and ordering of the original text. Can use simple techniques to start, maintain or end a short conversation. Can tell a story or describe something in a simple list of points.

PTE Academic and IELTS

PTE Academic Score	IELTS Score
N/A	9.0
89–90	8.5
84–88	8.0
76–83	7.5
66–75	7.0
56–65	6.5
46–55	6.0
36–45	5.5
29–35	5.0
23–28	4.5
10–22	No data

PTE Academic and TOEFL

PTE Academic Score	TOEFL iBT Score
85–90	No data
84	120
83	119
82	118
81	117
80	115–116
79	114
78	113
77	112
76	110–111
75	109
74	107–108
73	106
72	105
71	103–104
70	102
69	101
68	99–100
67	98
66	97
65	95–96
64	94
63	93
62	91–92

PTE Academic Score	TOEFL iBT Score
61	90
60	89
59	87–88
58	86
57	85
56	83–84
55	82
54	81
53	79–80
52	78
51	76–77
50	74–75
49	72–73
48	70–71
47	67–69
46	65–66
45	63–64
44	60–62
43	57–59
42	54–56
41	52–53
40	48–51
39	45–47
38	40–44
10–37	No data

Error of measurement

Tests aim to provide a measure of ability. PTE Academic measures the ability to use English in academic settings. Naturally, measures of a test taker's English language abilities will vary; some candidates will have higher scores than others. The degree to which scores among test takers vary is the 'score variance'. The purpose of testing is to measure 'true variance' in ability among students, but all measurement contains some error.

The degree to which the score variance is due to error is called the 'error of measurement'. The remainder of the variance is due to 'true variance' in ability among test takers. The error of measurement is related to the reliability of the test: a smaller measurement error means higher reliability of test scores.

The error of measurement can be interpreted as follows:

- The true score of a test taker is within a range of scores around the reported score.
- The size of that range is defined by the error of measurement. For example, if the reported score is 60 and the error of measurement is 3, then the true score, with 68% certainty, is within one measurement error from the reported score; that is within the range of 57 (60-3) and 63 (60+3).
- The true score, with 95% certainty, is within twice the measurement error; that is within the range of 54 (602x3) to 66 (60+2x3).

There are two main approaches to estimating the error of measurement. In Classical Test Theory (CTT) the reliability estimate is assumed to apply to any score on a test, irrespective of whether the score is low, medium or high. Therefore, the error of measurement is assumed to be the same size anywhere on the test's score scale. That is why in CTT we speak of the Standard Error of Measurement (SEM).

An alternative approach to estimating the error of measurement is used in modern test theory, commonly referred to as Item Response Theory (IRT). IRT recognizes that the reliability of a test is not uniform across an entire score scale. Tests tend to be less reliable towards the extreme low and high score ranges. Consequently, the size of the error of measurement tends to be larger towards these extreme scores. The size of the error is therefore conditional on the score and so, in IRT, we speak of Conditional Errors of Measurement (CEM).

The table below shows the average size of the CEM at five levels (A2 to C2) on the CEFR for the overall score and for the communicative skills scores that are provided on the PTE Academic score report. The size of the error is estimated by

averaging PTE Academic scores within each CEFR level for the full year of testing data in 2019.

PTE Academic Scores		Average Measurement Error				
		A2	B1	B2	C1	C2
Overall		2.8	2.8	3.2	3.8	4.6
Communicative skills	Listening	4.0	4.0	4.5	5.3	6.8
	Reading	4.2	4.1	4.8	5.8	7.2
	Speaking	4.6	4.7	5.6	6.9	9.6
	Writing	4.2	4.1	4.7	5.5	6.8

Measurement error for overall score and communicative skills scores at levels A2 to C2.

Test reliability

Directly related to measurement error is test reliability, which is another way of expressing the likelihood that test results will be the same when a test is taken again under the same conditions, and therefore how accurately a reported test score reflects the true ability of the test taker.

Reliability is expressed as a number between 0 and 1, where 0 means no reliability at all and 1 means perfectly reliable. For tests that are used to make important decisions, high reliability (0.90 or higher) is required. The following table provides the reliability estimates of the overall score based on testing data from 2021 and 2022.

For more information on the reliability of PTE Academic, refer to the paper *Establishing Construct and Concurrent Validity of Pearson Test of English Academic*, available at:

pearsonpte.com/research/published-research

Score	Overall
Reliability	0.95

Estimated reliability of overall score and communicative skills scores within PTE Academic score range of 53 to 79.

4. Automated scoring

As the worldwide leader in publishing and assessment for education, Pearson is using several of its proprietary, patented technologies to automatically score test takers' performance on PTE Academic. Academic institutions, corporations and government agencies around the world have selected Pearson's automated scoring technologies to measure the abilities of students, staff or applicants. Pearson customers using automated spoken and written assessments include eight of the 2008 Fortune Top 20 companies; 11 of the 2008 Top 15 Indian BPO companies; the U.S., German and Dutch governments; world sports organizations, such as FIFA (organizers of the World Cup) and the Asian Games; major airlines and aviation schools; and leading universities and language schools.

An extensive field test program was conducted to test PTE Academic's test items and evaluate their effectiveness as well as to obtain the data necessary to train the automated scoring engines to evaluate PTE Academic items. Test data was collected from more than **10,000 test takers** from 38 cities in 21 countries who participated in PTE Academic's field test. These test takers came from **158 different countries** and spoke **126 different native languages**, including (but not limited to) Cantonese, French, Gujarati, Hebrew, Hindi,

Indonesian, Japanese, Korean, Mandarin, Marathi, Polish, Spanish, Urdu, Vietnamese, Tamil, Telugu, Thai and Turkish. The data from the field test were used to train the automated scoring engines for both the written and spoken PTE Academic items.

By combining the power of a comprehensive field test, in-depth research and Pearson's proven, proprietary automated scoring technologies, PTE Academic fills a critical gap by providing a state-of-the-art test that accurately measures the English language speaking, listening, reading and writing abilities of a range of speakers.

Scoring written English skills

The written portion of PTE Academic is scored using the Intelligent Essay Assessor™ (IEA), an automated scoring tool that is powered by Pearson's state-of-the-art Knowledge Analysis Technologies™ (KAT™) engine. Based on more than 20 years of research and development, the KAT engine automatically evaluates the meaning of text by examining whole passages. The KAT engine evaluates writing as accurately as skilled human markers using a proprietary application of the mathematical approach known as Latent Semantic Analysis (LSA). Using LSA (an approach that generates semantic similarity of words

and passages by analyzing large bodies of relevant text) the KAT engine "understands" the meaning of text much the same as a human does.

IEA can be tuned to understand and evaluate text in any subject area, and includes built-in detectors for off-topic responses or other situations that may need to be referred to human readers. Research conducted by independent researchers as well as Pearson supports IEA's reliability for assessing knowledge and knowledge-based reasoning. IEA was developed more than a decade ago and has been used to evaluate millions of essays, from scoring student writing at elementary, secondary and university level, to assessing military leadership skills.

Scoring spoken English skills

The spoken portion of PTE Academic is automatically scored using Pearson's Versant technology. Versant technology is the result of years of research in speech recognition, statistical modelling, linguistics and testing theory. The technology uses a proprietary speech processing system that is specifically designed to analyze and automatically score speech from native and a range of linguistic backgrounds of English. In addition to recognizing words, the system locates and evaluates relevant segments, syllables and phrases in speech and then uses statistical modelling technologies to assess spoken performance.

To understand the way that the Versant technology is "taught" to score spoken language, think about a person being trained by an expert rater to score speech samples during interviews. First, the expert rater gives the trainee rater a list of things to listen for in the test taker's speech during the interview. Then the trainee observes the expert testing numerous test takers, and, after each interview, the expert shares with the trainee the score he or she gave the test taker and the characteristics of the performance that led to that score. Over several dozen interviews, the trainee's scores begin to look very similar to the expert rater's scores. Ultimately, one could predict the score the trainee would give a particular test taker based on the score that the expert gave.

This, in effect, is how the machine is trained to score, only instead of one expert teaching the trainee, there are many expert scorers feeding scores into the system for each response, and instead of a few dozen test takers, the system is trained on thousands of responses from hundreds of test takers. Furthermore, the machine does not need to be told what features of the speech are important; the relevant features and their relative contributions are statistically extracted from the massive set of data when the system is optimized to predict human scores.

Further information about automated scoring is available on our website:

pearsonpte.com/research/scoring

5. Scored Samples

The sections that follow show examples of how scoring operates in speaking and writing items. The scoring mechanisms in writing and speaking items are based on collecting data on multiple relevant traits in each item, giving them each a score and then converting them all to an overall score in either speaking or writing. The automated system is trained on the trait scores of hundreds of items scored by human expert markers. Once trained, our automated systems can then quickly score all new writing and speaking items quickly and accurately. The traits measured in PTE academic include:

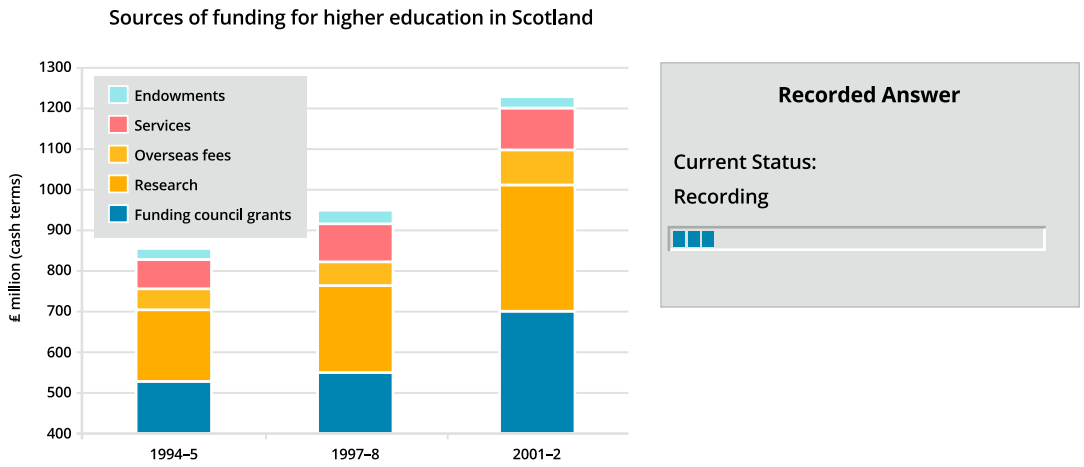
- Content
- Oral Fluency
- Pronunciation
- Form
- Development, structure and coherence
- Grammar
- General linguistic range
- Vocabulary

Spoken samples

The PTE Academic automated scoring system correlates highly with human ratings. Studies have been carried out to compare human and machine scores for the speaking item type ***Describe image*** using tasks such as the example below.

Example *Describe image* item

Look at the graph below. In 25 seconds, please speak into the microphone and describe in detail what the graph is showing. You will have 40 seconds to give your response.



Samples of test taker responses at B1, B2 and C1 were collected as well as comments from the Language Testing division of Pearson. The ratings on each response include a machine score and scores from at least two human markers. In cases where the two human rater scores differed, an adjudicator was used to provide a third human rating.

Scoring

The ***Describe image*** item is scored on three different traits:

Traits	Maximum raw score	Human rating	Machine score
Content	5	5	5
Oral fluency	5	5	5
Pronunciation	5	5	5
Maximum item score	**15**	**15**	**15**

These traits are scored as follows:

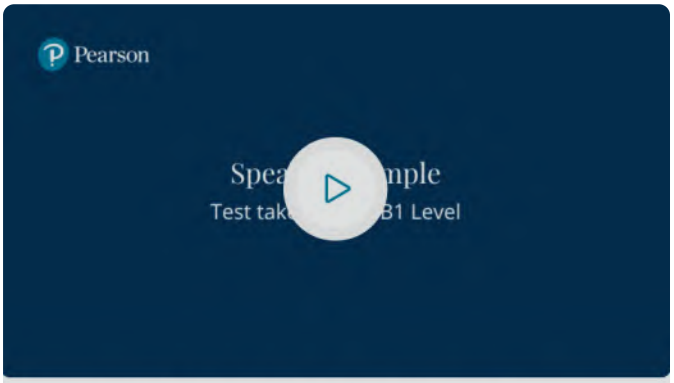
Content	Pronunciation	Oral fluency
5: Describes all elements of the image and their relationships, possible development and conclusion or implications.	**5 Highly proficient:** All vowels and consonants are produced in a manner that is easily understood by regular speakers of the language. The speaker uses assimilation and deletions appropriate to continuous speech. Stress is placed correctly in all words and sentence-level stress is fully appropriate.	**5 Highly proficient:** Speech shows smooth, rhythm and phrasing. There are no hesitations, repetitions, false starts or phonological simplifications.
4: Describes all the key elements of the image and their relations, referring to their implications or conclusions.	**4 Advanced:** Vowels and consonants are pronounced clearly and unambiguously. A few minor consonant, vowel or stress distortions do not affect intelligibility. All words are easily understandable. A few consonants or consonant sequences may be distorted. Stress is placed correctly on all common words, and sentence level stress is reasonable.	**4 Advanced:** Speech has an acceptable rhythm with appropriate phrasing and word emphasis. There is no more than one hesitation, one repetition or a false start. There are no significant phonological simplifications.

Content	Pronunciation	Oral fluency
3: Deals with most key elements of the image and refers to their implications or conclusions.	**3 Good:** Most vowels and consonants are pronounced correctly. Some consistent errors might make a few words unclear. A few consonants in certain contexts may be regularly distorted, omitted or mispronounced. Stress dependent vowel reduction may occur on a few words.	**3 Good:** Speech is at an acceptable speed, but may be uneven. There may be more than one hesitation, but most words are spoken in continuous phrases. There are few repetitions or false starts. There are no long pauses and speech does not sound staccato.
2: Deals with only one key element in the image and refers to an implication or conclusion. Shows basic understanding of several core elements of the image.	**2 Intermediate:** Some consonants and vowels are consistently mispronounced Some consonants and vowels are consistently mispronounced. At least 2/3 of speech is intelligible, but listeners might need to adjust to the accent. Some consonants are regularly omitted, and consonant sequences may be simplified. Stress may be placed incorrectly on some words or be unclear.	**2 Intermediate:** Speech may be uneven or staccato. Speech (if >= 6 words) has at least one smooth three-word run, and no more than two or three hesitations, repetitions or false starts. There may be one long pause, but not two or more.

Content	Pronunciation	Oral fluency
1: Describes some basic elements of the image, but does not make clear their interrelations or implications.	**1 Intrusive:** Many consonants and vowels are mispronounced, resulting in a strong intrusive foreign accent. Listeners may have difficulty understanding about 1/3 of the words. Many consonants may be distorted or omitted. Consonant sequences may be non-English. Stress is placed in a non- English manner; unstressed words may be reduced or omitted, and a few syllables added or missed.	**1 Limited:** Speech has irregular phrasing or sentence rhythm. Poor phrasing, staccato or syllabic timing, and/or multiple hesitations, repetitions, and/or false starts make spoken performance notably uneven or discontinuous. Long utterances may have one or two long pauses and inappropriate sentence-level word emphasis.
0: Mentions some disjointed elements of the presentation. May not deal properly with the prompt due to significant amounts of pre-prepared/memorized material.	**0 Non-English:** Pronunciation seems completely characteristic of another language. Many consonants and vowels are mispronounced, mis-ordered or omitted. Listeners may find more than 1/2 of the speech unintelligible. Stressed and unstressed syllables are realized in a non-English manner. Several words may have the wrong number of syllables.	**0 Disfluent:** Speech is slow and labored with little discernible phrase grouping, multiple hesitations, pauses, false starts, and/or major phonological simplifications. Most words are isolated, and there may be more than one long pause.

Test Taker responses

Test taker B1 Level



Listen to an example test taker response at B1 level.

Comment on response

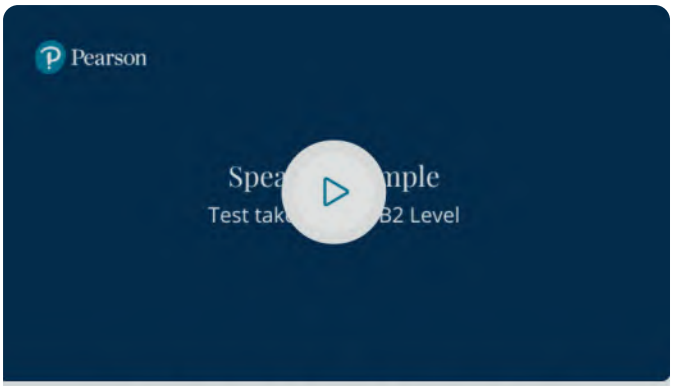
The response lacks some of the main contents. Only some obvious information from the graph is addressed. Numerous hesitations, pronunciation issues, poor language use and limited control of grammar structures at times make the response difficult to understand.

How the response was scored

The table below shows the machine scores and the human ratings that have been assigned to this response. The greyed out box indicates that the score is the same as the scores given by the first and second human rater.

Trait name	Maximum raw score	Machine score	Human rater 1	Human rater 2	Adjudicator
Content	5	1.69	2	2	2
Oral fluency	5	1.62	4	2	2
Pronunciation	5	1.41	2	2	2
Total item score	**15**	**4.72**	**8**	**6**	**6**

Test taker B2 Level



Listen to an example test taker response at B2 level.

Comment on response

The test taker discusses some aspects of the graph and the relationship between elements, though some key points have not been addressed. The rate of speech is acceptable.

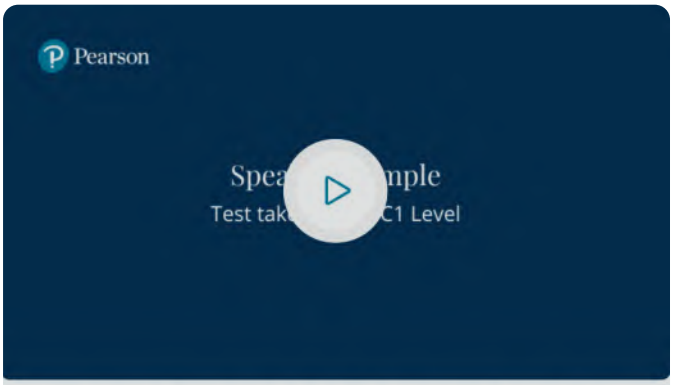
Language use and vocabulary range are quite weak. There are some obvious grammar errors and inappropriate stress and pronunciation.

How the response was scored

The table below shows the machine scores and the human ratings that have been assigned to this response.

Trait name	Maximum raw score	Machine score	Human rater 1	Human rater 2	Adjudicator
Content	5	2.50	2	3	2
Oral fluency	5	3.71	4	5	3
Pronunciation	5	3.28	3	4	2
Total item score	**15**	**9.49**	**9**	**12**	**7**

Test taker C1 Level



Listen to an example test taker response at C1 level.

Comment on response

The test taker discusses the major aspects of the graph and the relationship between elements. The response is spoken at a fluent rate and language use is appropriate. There are few grammatical errors in the response. The candidate demonstrates a wide range of vocabulary. Stress is appropriately placed.

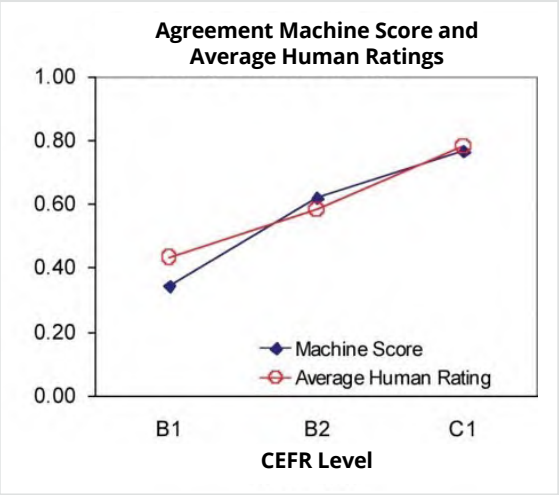
How the response was scored

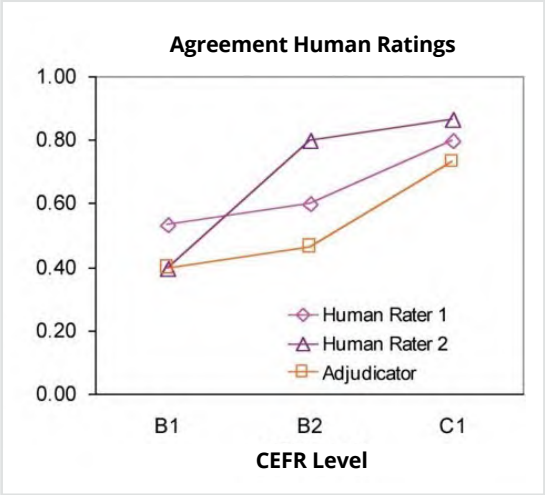
The table below shows the machine scores and the human ratings that have been assigned to this response.

Trait name	Maximum raw score	Machine score	Human rater 1	Human rater 2	Adjudicator
Content	5	2.70	3	4	3
Oral fluency	5	4.03	4	5	4
Pronunciation	5	4.02	5	4	4
Total item score	**15**	**10.75**	**12**	**13**	**11**

Overall performance rating

As shown from the scoring tables on the responses presented, the human ratings at trait level differed up to two score points out of six possible scoring categories (0 - 5). The two graphs below show the level of agreement of the total item score (sum of traits) of the human markers (graph on the left) and the agreement of the machine score with the average of the human ratings (graph on the right). The total item scores are rendered as a proportion of the total maximum item score (15) for the item. The human ratings vary substantially, especially for the B2 candidate, from a score that is only slightly higher than the score given to the B1 test taker, to a score that is close to the one given to the C1 test taker.





Note that these ratings were given by trained markers who had all recently passed a marker's exam. This example is therefore not typical for the human rating in general, but it shows that in some instances, especially for spoken responses, human markers have a hard time deciding on the most fitting score.

The automatic scoring system that has been trained on more than 100 human markers agrees quite well with the average human rating as shown in the graph on the right.

The machine-human comparison is part of the validation studies based on the field test responses for speaking, where 450,000 spoken responses were collected and scored, generating more than one million human ratings. The correlation between the human raw scores and the machine-generated scores for the overall measure of speaking was 0.89. In order to neutralize the effect of differences in severity amongst human ms, the human scores were scaled using Item Response Theory (IRT). The correlation with the machine scores then increases to 0.96. The reliability of the measure of speaking in PTE Academic is 0.91.

Score type	Human-human	Machine-human
Raw scores	0.87	0.89
IRT scaled	0.90	0.96

Written samples

The PTE Academic automated scoring system correlates highly with average human ratings. Studies were carried out to compare human and machine scores for the writing item type ***Write essay***, using tasks such as the example below.

Example *Write essay* item 'Tobacco'

You will have 20 minutes to plan, write and revise an essay about the topic below. Your response will be judged on how well you develop a position, organize your ideas, present supporting details, and control the elements of standard written English. You should write 200–300 words.

Tobacco mainly in the form of cigarettes, is one of the most widely-used drugs in the world. Over a billion adults legally smoke tobacco every day. The long term health costs are high - for smokers themselves, and for the wider community in terms of health care costs and lost productivity.

Do governments have a legitimate role to legislate to protect citizens from the harmful effects of their own decisions to smoke, or are such decisons up to the individual?

From the studies using these items, samples of test taker responses at B1, B2 and C1 are given as well as a comment from the Language Testing division of Pearson. Ratings on each response are provided including a machine score and scores from at least two human markers. In cases where the two human rater scores differed, an adjudicator was used to provide a third human rating.

Scoring

The item type ***Write essay*** is scored on 7 different traits:

Traits	Maximum raw score	Human rating	Machine score
Content	3	3	3
Form	2		
Development, structure and coherence	2	2	2
Grammar	2	2	2
General linguistic range	2	2	2
Vocabulary range	2	2	2
Spelling	2		
Maximum item score	**15**	**11**	**15**

The form and spelling traits do not require human ratings for training the automatic scoring systems as they can be objectively scored. It can be assumed (if the human markers work error-free) that the human rating on these two traits would have been identical to the machine score.

To make the total score from human rating comparable to the machine score, we need to take the score as a proportion of the maximum obtainable score by dividing the observed total score by the maximum possible score.

An item is not scored if the test taker's response does not meet the minimum requirements for the traits content and form (i.e., when a test taker scores 0 for content and/or form).

The traits are scored as follows:

Content	Form	Development, structure and coherence	Grammar
3: Adequately deals with the prompt.			
2: Deals with the prompt but does not deal with one minor aspect.	**2:** Length is between 200 and 300 words.	**2:** Shows good development and logical structure.	**2:** Shows consistent grammatical control of complex language. Errors are rare and difficult to spot.
1: Deals with the prompt but omits one major aspect or more than one minor aspect.	**1:** Length is between 120 and 199 or between 301 and 380 words.	**1:** Is incidentally less well structured, and some elements or paragraphs are poorly linked.	**1:** Shows a relatively high degree of grammatical control. No mistakes which would lead to misunderstandings.
0: Does not deal properly with the prompt. This includes responses that contain a significant amount of pre-prepared/ memorized material.	**0:** Length is less than 120 or more than 380 words. Essay is written in capital letters, contains no punctuation or only consists of bullet points or very short sentences.	**0:** Lacks coherence and mainly consists of lists or loose elements.	**0:** Contains mainly simple structures and/or several basic mistakes.

General linguistic range	Vocabulary range	Spelling
2: Exhibits mastery of a wide range of language to formulate thoughts precisely, give emphasis, differentiate and eliminate ambiguity. No sign that the test taker is restricted in what they want to communicate.	**2:** Good command of a broad lexical repertoire, idiomatic expressions and colloquialisms.	**2:** Correct spelling
1: Sufficient range of language to provide clear descriptions, express viewpoints and develop arguments.	**1:** Shows a good range of vocabulary for matters connected to general academic topics. Lexical shortcomings lead to circumlocution or some imprecision.	**1:** One spelling error
0: Contains mainly basic language and lacks precision.	**0:** Contains mainly basic vocabulary insufficient to deal with the topic at the required level.	**0:** More than one spelling error

Test Taker Responses

Test taker B1 Level

Tobacco, mainly in the form of cigarettes, is one of the most widely-used drugs in the world. Over a billion adults legally smoke tobacco everyday. Recently, it is not only the adult. Even the high school students or college students smoke just because they want to know how it feels. It is also not limited by gender. Lots of women are smokers. Even the old people still smoke, as if they do not care about their healthy. Become a smoker is like make someone just care about the good feeling of smoking and makes them to forget the risks they will face in the future. The long term health costs are high - for smokers themselves, and for the wider community in temrs of health care costs and lost productivity. The worst risk that the smokers will face is lung cancer, which can cause death. The governments have a legitimate role to legislate to protect citizens from the harmful effects of their own decisions to smoke. For example they make rule about no smoking area, in the street, and public place. But it also the decisions of each individual wheter they want to continue their life as a smoker and take all the risk, or stop and learn to life healthier. (211 words)

Comment on response

The response is a simple essay which gives a minimal answer to the prompt. The argument contains insufficient supporting ideas. The structure is lacking in logic and coherence. There is frequent misuse of grammar and vocabulary. Vocabulary range is limited and inappropriate at times.

How the response was scored

The table below shows the machine scores and the human ratings that have been assigned to this response. The greyed out box indicates that the score is the same as the scores given by the first and second human rater.

Trait name	Maximum raw score	Machine score	Human rater 1	Human rater 2	Adjudicator
Content	3	1.80	2	2	2
Development, structure and coherence	2	1.35	0	1	1
Form	2	2.00	n/a	n/a	n/a
General linguistic Range	2	1.03	1	1	1
Grammar	2	1.07	1	1	1
Spelling	2	0.00	n/a	n/a	n/a
Vocabulary range	2	0.93	1	2	1
Total item score	**15**	**8.18**	**5**	**7**	**6**

Test taker B2 Level

In my opinion it should be a combined effort of both government and an individual. In some countries specially in UK, government is tring to impose laws and regulations which discourage smoking, for example the law which prohibits smoking in pubs, bars and public areas. Also there are TV commercials and banners which explain the long term effects of smoking. As a result there has been some reduction in the number of people smoking before the law and now. But this effort is not enough. Uptil and unless an individual doesnt makes an effort himself the problem cannot be solved. One has to have control of his own body and will power to over come this habit turned necessity of the body. There has been a significant increase in amount of people who are approching mediacl practioners and NHS to help them to overcome this problem. There are also some NGO's who are working in this field. \n\nI think if we can spread awarness about the ill effects of smoking to teenagers, there will be less number of people who start smoking at the first place. It is a collective responsibilty of government and parents as well. To conclude i can say that youngsters are the people who get facinated by the whole idea of smoking, thus this concept should be changed by the efforts of government, media and by us as an individual. (234 words)

Comment on response

A systematic argument with appropriate highlighting of significant points and relevant supporting detail has been developed. Ability to evaluate different ideas or solutions to a problem has been demonstrated. However, some obvious grammar errors and inappropriate use of vocabulary can be found. There are also quite a number of spelling errors.

How the response was scored

The table below shows the machine scores and the human ratings that have been assigned to this response. The greyed out box indicates that the score is the same as the scores given by the first and second human rater.

Trait name	Maximum raw score	Machine score	Human rater 1	Human rater 2	Adjudicator
Content	3	2.25	3	1	2
Development, structure and coherence	2	1.17	2	1	2
Form	2	2.00	n/a	n/a	n/a
General linguistic Range	2	1.42	1	1	1
Grammar	2	1.68	1	2	3
Spelling	2	0.00	n/a	n/a	n/a
Vocabulary range	2	1.32	1	1	1
Total item score	**15**	**9.84**	**8**	**6**	**9**

Test taker C1 Level

Outlawing tobacco use would create unprecedented controversy. Billions of people worldwide smoke; whether they are chain smokers or recreational smokers. Also, there are several multi- million dollar cigarette companies that will also suffer many consequences if tobacco use is made illegal. We must also consider the thousands of employees who will be left unemployed if such a legislation is made. Unfortunately, it is an industry that makes ridiculous amounts of money for many people, so the likelihood of banning it is minimal.

Nonetheless, it is a change that would benefit society on many levels in the long run. Smoking causes so many health care issues, so if smoking is made illegal, morbidity and mortality rates would be reduced significantly. Quality of life will be improved dramatically, and it will allow more people to enjoy their lives significantly longer.

Legislators must also consider the rights of the individual. Should'nt every individual have the right to choose how they treat their body? The government can argue that these individuals may do as they wish, but then they must also suffer the consequences without government funding. They must take full responsibility for any health issues developed as a result of tobacco use, and not expect Medicare or health insurance to cover costs caused by their own irresponsible negligent decisions.

In essence, if individuals wish to make their own decisions to smoke, they must consider all the possible outcomes, and be willing to deal with these outcomes accordingly. (243 words)

Comment on response

Clear, well-structured exposition on the topic which touches upon the relevant issues. Points of view are given at some length with subsidiary points. Reasons and relevant examples are demonstrated. General linguistic range and vocabulary range are excellent. Phrasing and word choice are appropriate. There are very few grammar errors. Spelling is excellent.

How the response was scored

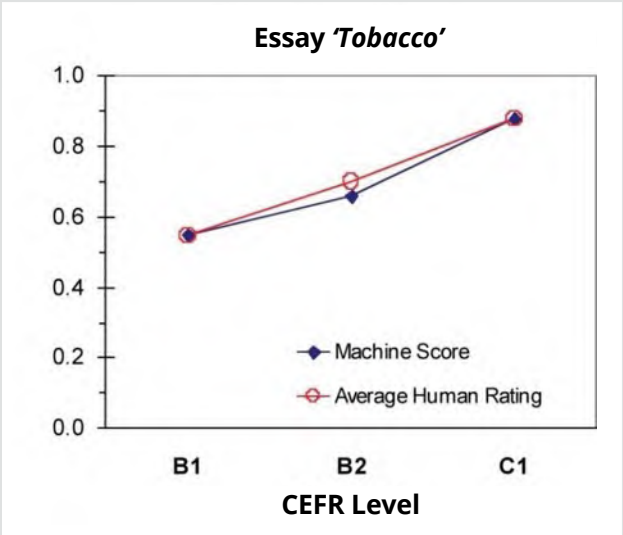
The table below shows the machine scores and the human ratings that have been assigned to this response. The greyed out box indicates that the score is the same as the scores given by the first and second human rater.

Trait name	Maximum raw score	Machine score	Human rater 1	Human rater 2	Adjudicator
Content	3	2.74	1	2	3
Development, structure and coherence	2	1.97	2	2	2
Form	2	2.00	n/a	n/a	n/a
General linguistic Range	2	2.00	2	2	2
Grammar	2	1.70	2	2	2
Spelling	2	1.00	n/a	n/a	n/a
Vocabulary range	2	1.82	1	2	2
Total item score	**15**	**13.23**	**8**	**10**	**11**

Overall performance rating

As can be seen from the scoring tables on the essay responses, the machine scores correspond closely to the average human score. Although there is some variation at the trait level, the total item scores agree to a high degree. To illustrate this agreement the graph below shows the machine scores and the average human scores.

The graph illustrates the total (proportional) item score from the machine and from the human ratings for the essay responses. The results show that the machine generated total item scores are closely aligned with the average over the human ratings.



The machine-human comparison is part of the validation studies based on the field test responses for writing, where 50,000 written responses were collected and scored, generating about 0.6 million human ratings.

The correlation between the human raw scores and the machine-generated scores for the overall measure of writing was 0.88. In order to neutralize the effect of differences in severity amongst human markers, the human scores were scaled using IRT. The correlation with the machine scores then increases to 0.93. The reliability of the measure of writing in PTE Academic is 0.89.

Score type	Human-human	Machine-human
Raw scores	0.87	0.88
IRT scaled	0.90	0.93

6. References

Using PTE Academic scores

American Council for the Teaching of Foreign Languages (1986) ACTFL Proficiency Guidelines. Hastings-on-Hudson, NY American Council for the Teaching of Foreign Languages (1999) ACTFL Proficiency Guidelines Speaking, (Revised), **https://www.actfl.org/sites/default/files/guidelines/ACTFLProficiencyGuidelines1999.pdf**

American Council for the Teaching of Foreign Languages (2012) ACTFL Proficiency Guidelines, **https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf**

Council of Europe (2001) Common European Framework of Reference for Languages: Learning, Teaching Assessment Cambridge: CUP

National Council of State Supervisors for Languages (2008) Linguafolio Self-Assessment Grid, **https://ncssfl.org/lfmodules/appendix-a/**

Concordance to other tests

Cassady, Jerrell C. (2001) Self-Reported GPA and SAT Scores. ERIC Digest. ERIC Identifier: ED458216

Council of Europe (2001) Common European Framework of Reference for Languages:

Learning, Teaching, Assessment. Cambridge: CUP

Clesham, R. & Hughes, S. (2020). **2020 Concordance Report PTE Academic and IELTS Academic**

ETS (2001) TOEFL Institutional Testing Program (ITP) and TOEIC Institutional Program (IP):

Two On-Site Testing Tools from ETS at a Glance. Handout Berlin Conference 2001. Princeton: Educational Testing Service

ETS (2005) TOEFL ® Internet-based test: Score comparison tables. Princeton: Educational Testing Service

Linacre, J.M (1988; 2005) A Computer Program for the Analysis of Multi-Faceted Data. Chicago, IL: Mesa Press

7. Glossary

ACTFL – American Council on the Teaching of Foreign Languages – An individual membership organization of language educators, students and administrators dedicated to the improvement of the teaching and learning of all languages at all level of instruction organization.

CEFR (also known as CEF) – The Common European Framework of Reference for Languages put together by the Council of Europe to standardize the levels of language exams in different regions. Other exams are mapped to the CEFR.

Communicative skills – Four skills for which PTE Academic test takers receive reported scores. These skills are listening, reading, speaking, and writing.

Concordance studies – The relationship between two or more scales of measurement.

Global Scale of English (GSE) – The Pearson GSE is a truly global English language standard. Based on research involving over 6000 teachers from more than 50 countries, it extends the Common European Framework of Reference (CEFR) by pinpointing on a scale from 10 to 90 what needs to be mastered for the four skills of listening, reading, speaking and writing within a CEFR level, using a more granular approach. For additional information, visit: **https://www.pearson.com/english/about/gse.html**

Integrated skills items – Items on the test that require the use of more than one skill such as assessing reading and speaking, listening and speaking, reading and writing, listening and writing, or listening and reading.

Intelligent Essay Assessor™ (IEA) – An automated scoring tool that is powered by Pearson's state-of-the-art Knowledge Analysis Technologies™ (KAT™) engine.

Item Response Theory (IRT) – A testing theory. IRT is based on the relationship between an individual's performance on a test item and that individual's levels of performance on an overall measure of the ability that item was designed to measure.

Error of measurement – The degree to which the score variance is due to error.

Formal aspects – The form of a response: for example, whether it is over or under the word limit for a particular item type.

IELTS – International English Language Testing System. This test measures the language proficiency of people who want to study or work where English is used as a language of communication.

LinguaFolio Self-Assessment Grid – An assessment tool that relates language levels to the scales of both the ACTFL (American Council on the Teaching of Foreign Languages) and the CEFR (Common European Framework of Reference for Languages).

National Council of State Supervisors for Languages (NCSSFL) – An organization of education agency personnel from across the United States who have the responsibility of foreign/world language education at the state level.

Versant technology – A proprietary speech processing system that is specifically designed to analyze and automatically score speech from a range of linguistic backgrounds.

Overall score – Score based on test taker's performance on all test items.

PTE Academic – Pearson Test of English Academic. PTE Academic is a 2-hour long, computer-based assessment of a person's English language ability in an academic context. The test assesses an individual's communicative skills of listening, reading, speaking, and writing through questions using authentically-sourced material.

Score variance – The degree to which scores among test takers vary.

TOEFL iBT – A test that measures the ability to use and understand English at the university level, and evaluates how well the test taker combines listening, reading, speaking, and writing skills to perform academic tasks.

Traits – Items measured in PTE Academic that contribute to overall scores. These include content; oral fluency; pronunciation; form; development, structure and coherence; grammar; general linguistic range; and vocabulary.

Appendix

The following table shows the PTE Academic score aligned to the CEFR discussed in section 3 and describes performance indicators at these levels.

PTE Academic Score	Common European Framework Level	Level Descriptor[1]	What does this mean for a score user?
85 - 90	**C2**	Can understand with ease virtually everything heard or read. Can summarize 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.	C2 is a highly proficient level and a student at this level would be extremely comfortable engaging in academic activities at all levels.
76 - 84	**C1**	Can understand a wide range of demanding, longer texts and recognize 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 organizational patterns, connectors and cohesive devices.	C1 is a level at which a student can comfortably participate in all post-graduate activities including teaching. It is not required for students entering university at undergraduate level. Most international students who enter university at a B2 level would acquire a level close to or at C1 after living in the country for several years, and actively participating in all language activities encountered at university.

[1] © The copyright of the level descriptors reproduced in this document belongs to the Council of Europe.

PTE Academic Score	Common European Framework Level	Level Descriptor[1]	What does this mean for a score user?
59 - 75	**B2**	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization. Can interact with a degree of oral fluency and spontaneity that makes regular interaction with other 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 disadvantages of various options.	B2 was designed as the level required to participate independently in higher level language interaction. It is typically the level required to be able to follow academic level instruction and to participate in academic education, including both coursework and student life.
43 - 58	**B1**	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 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.	B1 is insufficient for full academic level participation in language activities. A student at this level could 'get by' in everyday situations independently. To be successful in communication in university settings, additional English language courses are required.

PTE Academic Score	Common European Framework Level	Level Descriptor[1]	What does this mean for a score user?
30 - 42	**A2**	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.	A2 is an insufficient level for academic level participation.
10 - 29	**A1 or below**	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.	A1 is an insufficient level for academic level participation.