Pearson
Test of English

Dr. Rose Clesham and Sarah R. Hughes

The Enhanced PTE Academic

Test Description and Validation Report

Contents

Pearson

Background

Pearson Test of English (PTE) Academic is a computer-based international test of English as a second language.

PTE Academic was launched in 2009, in response to demand from higher education, governments, and professional bodies for a test that would securely and accurately measure the English communication skills of international students and economic migrants in academic and professional contexts. The purpose of this test is to measure test takers' English language competency in listening, reading, speaking and writing for academic and skilled migration purposes.

The constructs measured are the communicative language skills needed for reception, production, and interaction in both oral and written modes, as these skills are necessary to successfully follow academic courses and to actively participate in the targeted professional language environments.

The foundation of PTE Academic was and is the Council of Europe Framework of Reference for Languages (CEFR, Council of Europe, 2001).

PTE Academic was aligned to the CEFR from inception, using a fine-grained proficiency scale of 10–90, calibrated against the six CEFR levels. Whereas the CEFR describes attainment in six broad levels from A1 (low basic) to C2 (fully proficient), the PTE Academic test uses a standardized, numerical scale from 10–90, so that learners can see the step-by-step progress they are making and maintain motivation. It reports a 10–90 score on the four communicative skills (listening, reading, speaking, and writing), and an overall test score on the same scale. The overall score is based on performance on all scored questions. Each question score contributes to the overall score.

> "PTE Academic was aligned to the CEFR from inception, using a fine-grained proficiency scale of 10–90, calibrated against the six CEFR levels.

Introduction

Research, expansion and elaboration of the CEFR has continued since the original 2001 publication, and it has undergone several key updates since its original publication, with the latest Companion volume published in 2020 (Council of Europe, 2020).

Key updates included:

- Expanded Descriptors: To include areas like online interaction and mediation.
- New Scales: New scales have been added for mediating text, mediating concepts, and mediating communication.
- Updated Levels: The descriptors for the A1 and C levels have been revised to provide clearer and more detailed guidance.

It has been an ongoing focus of PTE research to develop new question types and scoring rubrics that complement and expand the linguistic competencies of the test in line with CEFR enhancements.

It is important to note that the enhanced PTE Academic test in large part retains its essential structure and standards:

- No question types have been removed
- It tests all four language skills using a combination of single and integrated skills
- Most of the scoring rubrics are unchanged
- The test retains its score alignment to the CEFR

The main revisions have been the introduction of two new extended speaking question types, changes to the frequency of some question types in each test, and enhanced scoring procedures and scoring rubrics for extended speaking and writing.

This paper outlines the new enhanced PTE Academic test structure, the research underpinning the changes, the development and rationale of the two new PTE Academic speaking questions and scoring enhancements.

The Enhanced PTE Academic Test Structure

The following table shows the changes to the test structure and compares the number of question types between the current and updated PTE Academic test.

Total number of questions in test			
Section	**Question Type**	**Current PTE**	**PTE 2025 revisions**
Part 1: Speaking & Writing	Read Aloud	6	6
	Repeat Sentence	10	10
	Describe Image	3	5
	Retell Lecture	1	2
	Answer Short Question	5	5
	Respond to a Situation	0	2
	Summarize Group Discussion	0	2
	Summarize Written Text	1	2
	Write Essay	1	1
Part 2: Reading	Fill in the Blanks (Dropdown)	5	5
	Multiple Choice, Multiple Answers	1	2
	Reorder Paragraph	2	2
	Fill in the Blanks (Drag and Drop)	4	4
	Multiple Choice, Single Answer	1	2
Part 3: Listening	Summarize Spoken Text	1	1
	Multiple Choice, Multiple Answers	1	2
	Fill in the Blanks (Type In)	2	2
	Highlight Correct Summary	1	2
	Multiple Choice, Single Answer	1	2
	Select Missing Word	1	1
	Highlight Incorrect Words	2	2
	Write from Dictation	3	3
TOTAL		**52**	**65**

The updated test is composed of approximately 65 scored questions across 22 question types. These question types are organized into the same three sections as the 52-question test. The test still takes approximately 2 hours to complete.

Updates to the PTE Academic test focus on increasing the opportunities for test takers to produce extended responses for speaking and writing. There is a greater emphasis on spontaneous production in response to novel stimuli in the Describe Image and Summarize Spoken Text question types in speaking and the Summarize Written Text question type in writing. There are also small increases in the number of listening and reading constrained questions and a reduction of three integrated skills questions, now being reallocated to solely one skill; the Answer Short Question, Fill in the Blanks and Read Aloud questions. These changes have been made to improve clarity and precision of the assessment of all four skills.

As shown in the table above, Pearson has introduced two additional question types. These questions in particular show how PTE Academic is developing in line with key enhancements of the CEFR, with a greater focus on responses that involve interaction and mediation.

Of the 22 different question types, 8 are classed as integrated, meaning they assess more than one language skill. For example, a task might ask the test taker to read a text and write a short summary or listen to a question and give a short spoken response. Integrated question types bring a number of advantages to the assessment process. By collecting data for multiple skills in the same task, a more complete, authentic assessment of a test taker's proficiency can be made in a shorter period of time. Integrated question types reflect the real-life skills test takers will need in a higher-education, social and work setting, improving test validity and authenticity. They also enable the assessment of higher order thinking skills, which require test takers to respond in real time to written or spoken input materials at length and in their own words. This provides a level of authenticity not possible in tests that focus on the assessment of skills in isolation.

Overall, the updates to the PTE Academic test specification were designed to ensure that the test as a whole provides more opportunities for test takers to demonstrate a range of essential language skills and competencies. The test retains all the qualities of the 52-question test and remains a fast, fair and reliable measure of language proficiency.

Information about the design of the two additional question types, Respond to a Situation and Summarize Group Discussion, is provided below, along with the rationale for their inclusion in the test and explanation of their contribution to the assessed construct.

Speaking Task Enhancements

The speaking sections in PTE Academic uses various constructed-response and open-response formats, requiring test takers to produce authentic samples of speech. The elicited responses provide opportunities to assess test takers' linguistic, sociolinguistic, discourse, psycholinguistic, and functional competencies.

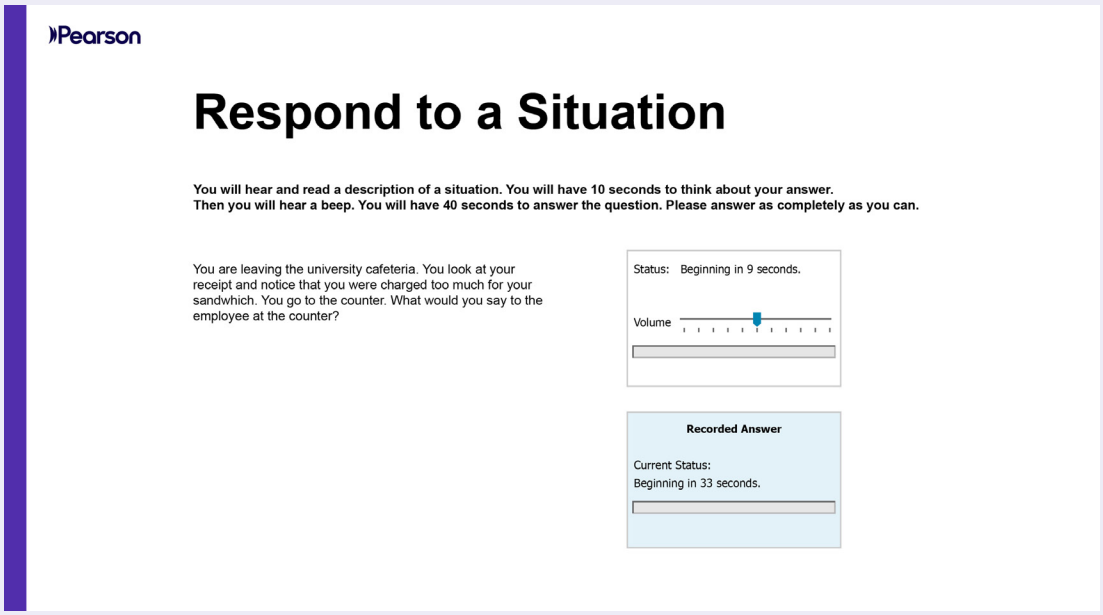
PTE Academic is a computer-based and computer-scored test and as such does not measure speaking ability in a face-to-face situation with an examiner. This does not mean, however, that the skills required to take part in face-to-face conversations are not assessed in PTE Academic. Task types have been designed based on research into the constituent skills required in face-to-face situations in academic and professional contexts, from lower-order skills such as pronunciation to higher-order processes such as discourse construction.

As discussed above, and in line with Pearson's commitment to continuous improvement, our research has been focused on developing new question types and scoring rubrics that complement and expand the linguistic competencies in line with CEFR recommendations and enhancements, maximizing the construct and face validity of the test and encouraging more spontaneous speaking and engagement with novel stimuli.

The Two New Question Types

1. Respond to a Situation

This question type provides opportunities for extended spoken responses and higher order skills relevant to academic and professional domains. The image below shows an example of the question type as it appears to test takers.



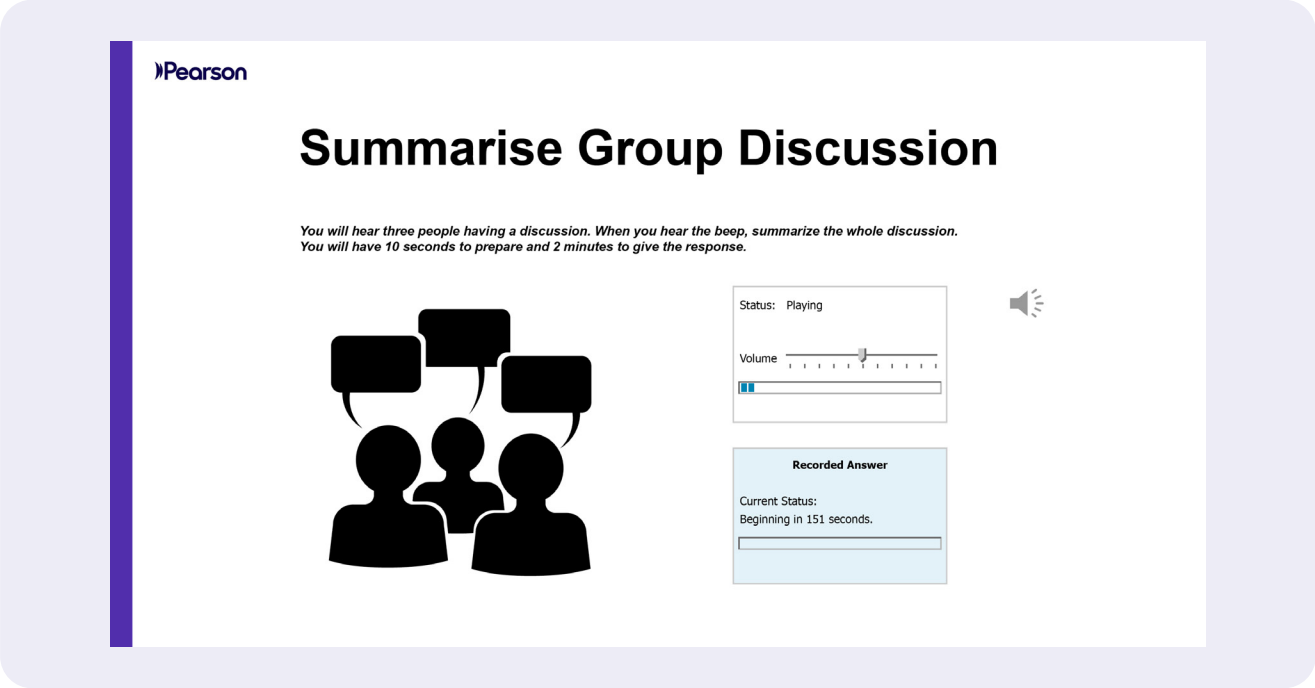
In this speaking task, test takers listen to and read a brief scenario and are then asked to respond orally as if they were in the situation. Test takers have 10 seconds to prepare a response and 40 seconds to give their response to each situation.

Test takers are expected to give relevant, appropriate responses in a clear and coherent manner. They are scored on the appropriateness of their response to the given situation, as well as the fluency and pronunciation of their speech.

Interactional competence is an integral part of academic and professional language use. This includes the concept of pragmatics, which ties together language structures (syntax, lexis, and phonology) and the contexts in which they are used in the real world. The Respond to a Situation task elicits aspects of pragmatic ability in a relatively open, long-turn response, simulating real world interaction. The inclusion of this kind of task adds breadth to the spoken English proficiency construct measured by PTE Academic, as well as encouraging the use of higher order skills related to improvisation and simulated interaction. Test takers respond as if participating in the scenario, and the response needs to take into consideration the specifics of each situation. Scenarios are drawn from a diverse range of contexts, functional purposes, levels of formality, and communicative intentions, ensuring that the tasks are engaging, non-routine, and relevant for both academic and professional test takers.

2. Summarize Group Discussion

This question type also includes more opportunities for extended spoken responses, and the use of higher order skills relevant to academic and professional domains. The image below shows an illustrative example of the question type.



In this speaking/listening integrated skills task, test takers listen to a discussion among three speakers and then retell what was discussed in their own words. Each discussion is approximately 2 to 3 minutes in length and test takers are then given up to 2 minutes to complete the retelling task.

This question type is scored on the content of the response, as well as the fluency and pronunciation of the delivery. Test takers are expected to retell the discussion, including the main points and important details, and provide an accurate summary. In order to do this, test takers must understand the content in order to fully and accurately produce a coherent constructed spoken response. The listening texts include a variety of authentic features that would be expected in real world discussions, such as different accents, fillers and hesitations, false starts, and self-corrections. They take place in real time, meaning the texts are only played once, so test takers must process language as they listen, just as they would do in real world situations.

Supporting Research

The two new speaking question types were designed and developed to extend the form and construct of the speaking section of the test by assessing mediation and interaction in different forms.

This is in line with the recommendation that adequate coverage of the target language use domain is essential for valid interpretations about test taker abilities in the second language (Bachman & Palmer, 1996).

One of the new question types (Respond to a Situation) focuses on the improvisational monologic skill of responding to a given situation (meaning that the text is delivered by a single speaker).

The other (Summarize Group Discussion) deals with the demand of listening to a lengthy dialogic discussion (between multiple speakers) and giving an extended accurate summary.

In comparing monologic and dialogic speaking test tasks, Papageorgiou, Stevens, and Goodwin (2012) found that there were substantial differences in the outcomes of these two kinds of tasks, suggesting that it would be fruitful to include both discourse types in language tests. Both monologic (single speaker) and dialogic (conversational) modes of discourse are common in the target language use domains for academic study and professional migration and, therefore, usefully represented in high-stakes assessments that target these domains.

> “Rather than a linear argument, this more informal mode of interpersonal discourse requires the construction of a mental representation of what each speaker has said and how that relates to what other speakers have said in the sequence of the conversation.

The Respond to a Situation task elicits aspects of interactional competence in a relatively open, long-turn response. The focus is on the test taker’s spoken response, with the scenario presented in both written and audio text, offering test takers a choice of accessing the scenario through whichever receptive skill they prefer. Responses are scored on the appropriacy of the response to the given situation, as well as the fluency and pronunciation of their speech.

Interactional competence is an integral part of the language construct and has been shown to overlap but also differ from more conventional measures of second language speaking ability (Roever & Ikeda, 2022). This competence includes the concept of pragmatics, which ties together language structures (syntax, lexis, and phonology) and the contexts in which they occur in the real world.

The Respond to a Situation task was adapted from an early effort by Hudson, Detmer, and Brown (1995) which was shown to yield high reliability values (.75-.86) for similar kinds of task types. This task was designed to add breadth to the second language assessed construct as well as providing opportunities to assess higher order skills related to improvising in simulated interaction in English.

The Summarize Group Discussion task has been designed to be complementary, but different to the Retell Lecture question type, in which test takers summarize only one speaker's point of view. Lectures used in Retell Lecture may involve complex subject matter, but the information is monologic or presented by only a single speaker. The Summarize Group Discussion task, on the other hand, involves content, which is more conversational, with elaboration coming from the dialogic interaction between three different speakers, the different ideas presented, and the ways in which the speakers' identities, roles, and attitudes towards each other are made evident during the discussion. Rather than a linear argument, this more informal mode of interpersonal discourse requires the construction of a mental representation of what each speaker has said and how that relates to what other speakers have said in the sequence of the conversation.

Revised Scoring Rubrics

The majority of scoring rubrics have not been changed in the enhanced PTE Academic test.

The automated scoring system continues to score traits that benefit from the objective and unbiased measurement of aspects of spoken and written responses, such as grammatical accuracy, pronunciation, vocabulary range, and fluency. The strength of the automated scoring system is its ability to measure such traits with precision and reliability across a diverse range of test takers with different accents, genders, and levels of proficiency.

As the new question types were being developed, the accompanying Content scoring rubrics were also a key focus area. These needed to differentiate test takers across the full range of language abilities in academic and skilled professional contexts and be as detailed and nuanced as possible.

> “ The finalized scoring rubrics for the new question types were the result of the underpinning research, listening to stakeholder feedback, and also involving senior expert human raters to trial, review, revise and finalize them throughout piloting and field testing.

The process began by aligning the draft scoring rubrics to the CEFR descriptors (Council of Europe, 2001), as well as to scoring rubrics from other high-stakes assessments targeting the same cohort of test takers. The goal was to ensure that the new question type rubrics captured the breadth and depth of the intended language proficiencies being assessed by those tasks.

Additionally, Pearson also conducted Test Review Groups (TRGs) with professional organizations and universities to better understand the construct of English proficiency valued by academic and professional stakeholders (Clesham et al., 2023). The findings from these review groups provided vital evidence related to target language use (TLU) domains and fed into the development of the new scoring rubrics.

For each question type, the facets of the intended construct were described. In the case of the new question types, the new Content score for the Responding to a Situation question type is evaluated based on the success of the response in achieving the primary goal of the communication (e.g. apology, request etc) while taking into account the context of the situation given in the prompt. The Content scoring of this task considers how effectively the situation is addressed, how well the response considers all elements of the context provided in the prompt, how effectively and flexibly the speaker communicates, how situationally appropriate and varied the language use is, and the level of restriction evident in the speaker's ability to express ideas.

The Summarize Group Discussion question type is intended to assess listening comprehension of multiple perspectives in a discussion, the ability to synthesize information, express ideas precisely, organize ideas logically, and communicate all of this in spoken English. The Content rubric for the Summarize Group Discussion question type is therefore designed to evaluate completeness and accuracy, synthesis, precision of expression, and logical organization of the response.

The finalized scoring rubrics for the new question types were the result of the underpinning research, listening to stakeholder feedback, and also involving senior expert human raters to trial, review, revise, and finalize them throughout piloting and field testing.

The new scoring rubrics are based on a rating scale of 0 to 6. The number of levels was guided by the dual priorities of being able to differentiate between proficiency levels as

> "The Write Essay task assesses more traits than the other extended question types and as a result, Pearson decided to extend the rubrics of three of the nuanced writing traits.

accurately as possible whilst at the same time having a scale that human raters could practically use and arrive at the same ratings as other human raters. This final decision was informed by expert panels and prior internal standard setting research that suggested 6 was the optimal number of levels across which raters could differentiate performance with high levels of accuracy and reliability. The rating scales are not intended to map directly to CEFR levels (i.e. the score of 6 is not intended to represent C2), however the rating scales are intended to represent the full span of ability from A1 to C1/C2.

As the research and development of the updated test progressed, Pearson also used the research development as an opportunity to update the Content rubrics of all the other extended response tasks in speaking and writing.

As for the new question types, the number of levels of performance described by the rating scales was guided by the dual priorities to both differentiate precisely and to produce reliable scores fit for high-stakes decision-making.

Therefore, as well as applying to the new question types, the revised 6-point scoring rubric is also applied to existing speaking tasks Describe Image and Lecture Retell. A Content rubric from 0-4 is applied to the Summarize Spoken Text task.

The Write Essay task assesses more traits than the other extended question types and as a result, Pearson decided to extend the rubrics of three of the nuanced writing traits. Scoring rubrics of 0–6 are applied to the following traits for this question type: Content, Development Structure and Coherence and General Linguistic Range. A Content rubric from 0–4 is applied to the Summarize Written Text task. Full details on the updated extended question rubrics can be found in the updated PTE Academic Score Guide.

Ensuring Test Integrity and Increasing Human Oversight

A large benefit of PTE Academic is the quick turnaround time for providing test results.

This is possible partly because of the automated scoring procedure in place for all closed and short response question types. Some question types, such as multiple choice questions, are simple to automatically score because the answers are fixed and predetermined. Other question types are more complicated to score because they require test takers to produce open responses of varying lengths. The automated scoring system utilizes AI speech recognition and natural language processing research. The scoring system has also been developed using machine learning techniques which ensure that the computer grades spoken and written responses in a way that replicates human examiners. The process of machine learning involves human experts at each step of the journey, from the definition of assessment rubrics to the training of the computer algorithms based on the responses of thousands of test takers that have themselves been graded by human assessors. Human judgment has always been incorporated throughout the scoring process to ensure the quality and integrity of scores.

> “ It is the responsibility of testing organizations like Pearson to ensure that tests are as valid, reliable, and fair as possible, and that they are regularly monitored in terms of both the qualities of the tests themselves and any emergent behaviors that encourage negative washback in terms of test preparation and test-taking strategies.

As part of Pearson’s ongoing research into test preparation and test integrity, a large-scale research programme has been undertaken to use Pearson’s automated systems to identify potentially gamed or templated test responses, and then make this information available to human raters when they are scoring responses, to help them determine whether the response is legitimate or if it should not be considered for further scoring.

Most test takers prepare for and take high-stakes English language tests in good faith and it is therefore important that test takers and stakeholders have trust in the operation and standards of the assessment system. It is the responsibility of testing organizations

like Pearson to ensure that tests are as valid, reliable, and fair as possible, and that they are regularly monitored in terms of both the qualities of the tests themselves and any emergent behaviors that encourage negative washback in terms of test preparation and test-taking strategies. (see Hughes and Clesham, 2024).

As well as dealing with the increasing threats of templating and negative washback behaviors, the duality of technology and human judgment can also provide a baseline for evidencing the accuracy and reliability of Pearson's automated scoring systems and ensure that public trust and confidence in high-stakes assessment is merited.

To reinforce public trust and confidence in PTE Academic as a valid and reliable test, Pearson has made some changes to the way in which open responses are scored, incorporating the use of hybrid human and automated scoring for all extended speaking and writing tasks in the enhanced PTE Academic test. The Content trait in all extended speaking and writing tasks is now scored by the computer and human raters. For the Write Essay question type, the traits of Development, Structure and Coherence, and General Linguistic Range now also incorporate human and automatic scoring.

The emergence of technology has transformed the delivery, accuracy, and fairness of global English language testing. It is therefore logical that the use of technology is also increasingly required to protect test integrity. The development of automated monitoring systems that can flag gamed spoken and written responses which can then also be scored by human expert raters utilizes both the power and reach of technology, paired with the essential element of human judgment.

Therefore, although Pearson is at the forefront in terms of utilizing new technologies to score and to maintain test integrity, we also want to ensure that active human expert judgment is at the heart of our systems, combining the best of what automated systems can offer with the re-assurance of nuanced human judgment.

The human-scored and machine-scored trait scores for all extended speaking and writing tasks are combined into the final score for the response. The automated scoring system will continue to score traits that benefit from objective and unbiased measurement, such as grammatical accuracy, pronunciation clarity, vocabulary range and fluency. The strength of the automated scoring system is its ability to measure such traits with precision and reliability across a diverse range of test takers with different accents, genders, and levels of proficiency.

In this way, the enhanced PTE Academic ensures that every extended response is evaluated using the most appropriate and robust scoring methodologies, combining the best of human and automated scoring.

Field Testing

In order to validate the new question types within a revised test form and also evaluate the technical functioning of the test, Pearson conducted a large-scale field test of the enhanced PTE Academic test.

To recap, the enhanced PTE Academic test differs from the current PTE Academic test in the following ways:

- The addition of two new speaking question types, adding levels of mediation and interaction to the test in terms of assessing pragmatic communicative competence and interpreting multiple speaker dialogues.
- The introduction of expanded content scoring rubrics for all extended spoken and written responses to maximize construct and face validity and differentiate the scoring of intended skills across academic and skilled professional contexts.
- An increase in the number of tasks to 65, with adjustments to the proportions of question types in the test to ensure balance and skill attributions.
- The introduction of additional human oversight and enhanced automated scoring systems to further strengthen test integrity.

Although the new question types had been previously trialled in research, the field test ensured that they were trialled with a large, representative sample of PTE Academic test takers and in the context of the other question types on the test. In the updated test, human raters score certain traits for extended constructed responses alongside the automated scoring system. For the field test, these traits were scored exclusively by human raters as a first step to retraining all automated scoring models for these traits.

Recruitment for the PTE Academic field test mainly took place in India, China, Australia, and the United Kingdom, as these countries account for the majority of PTE Academic test takers.

2,879 participants completed the field test. All field test takers sat the enhanced PTE Academic test, taken under controlled conditions in official PTE test centers around the world. Test takers were briefed on the new test structure and provided with information regarding the new question types, including examples, to ensure they were prepared and able to complete the field test in good faith and with full effort.

The new question types were well received by test takers in terms of the authenticity of the tasks presented and results indicated that test takers across the ability range could engage with and achieve scores on the new tasks.

The tasks differentiated proficiency levels well, elicited good samples of extended speech from test takers, assessing the intended targeted skills and allowing for higher order language skills to be demonstrated at the top end of language proficiency.

The finalization of the scoring and rating scales rubrics was informed by input from experienced senior human raters, panels of stakeholders representing academic and professional domains, reviews from the PTE Technical Advisory Group and external experts, alongside empirical data from the large-scale field test.

The field-test data demonstrated that the revised rubrics worked as intended, ie. that all score points were used by raters and showed clear differentiation across the ability range on the intended domain-relevant skills.

The resulting reliability statistics also demonstrated a high level of consistency, with inter-rater reliability statistics and test reliability statistics exceeding .80.

Feedback from Test Takers

As part of the validation of the enhanced PTE Academic test, feedback from the concordance study participants was collected. Details on the concordance study are given in the next section. 663 survey responses were collected, followed by 20 in-depth interviews.

The topics of the surveys and interviews included test taker background, feedback on the test in general and on the new question types, test familiarity, and their attitudes to the use of AI in high-stakes testing. The full report, *Test Taker Perceptions of the Enhanced PTE Academic Test*, has a forthcoming release. The following quotes are examples of test taker views on the two new speaking tasks:

Respond to a Situation:

"I like this kind of question because it's very creative and it's a simple scenario with which people can have anywhere."

"I will say it's extremely relevant. It's something it's a conversation that may come up in daily life, yes."

"Maybe you are at the intersection on a main highway, and you are taking right turn. Your friend is suggesting not to take right turn, but you are thinking that you should take the right turn."

Summarize Group Discussion:

"I think it is pretty much relevant to like what the English tests should be. Because in day-to-day life we come across such situation like that people are talking like for an example, like I'm working and I'm, I have two colleagues, we are sharing some important information regarding our work."

"For example, when I live with my housemates, we usually speak in a group and share our thoughts over something. It will be a scenario kind of like this."

"In my second year I joined a few student clubs and sometimes a lot of university professors are also part of those clubs. And yeah, we do have a lot of discussions on what could be good for like open days like when we set up our stalls and what wouldn't work out as well."

Alignment to the CEFR

As global English language tests each have their own score reporting scale, the CEFR has served as a central and necessary frame of reference to support score interpretation and use.

PTE Academic differs from most tests and exams in that the test was designed to measure language competence according to the principles of the CEFR from inception. The original development process of the PTE Academic test included the stages of Familiarization, Specification, and Standardization, as indicated by the CEFR (Council of Europe, 2009), culminating in a robust alignment of the scores of the PTE Academic test and the CEFR levels (Zheng & De Jong, 2011).

It is a key responsibility of testing organizations to establish and validate test alignments to language frameworks such as the CEFR, however it is also important to maintain and update them on an ongoing basis. Over time, various factors may cause these alignments to shift, including changes in test familiarity, testing populations, or test design. Such changes should be monitored, and alignments periodically revalidated or updated.

As Pearson carried out the research and development to enhance the PTE Academic test, it was an essential element to consider how any revisions might impact the established PTE Academic-CEFR alignment. It was decided from the outset that the standards of the test would be maintained in terms of the alignment with the CEFR.

Following guidance from Aligning Language Education with the CEFR: A Handbook, (Council of Europe, 2022), Pearson used both standard setting and statistical linking procedures to triangulate performance and standards outcomes for the enhanced test. Steps included expert standard setting panels carrying out modified Angoff activities to judge task difficulties in relation to CEFR levels, examinee-centered Body of Work activities to confirm test-taker threshold performance alignment between the recalibrated enhanced PTE Academic test and CEFR levels, and statistical equating analyses to ensure comparable standards between the enhanced PTE Academic test and the current version of the test. The goal of these analyses was to ensure continuity of score meaning for the enhanced PTE Academic test.

“ As Pearson carried out the research and development to enhance the PTE Academic test, it was an essential element to consider how any revisions might impact the established PTE Academic-CEFR alignment.

Concordance with IELTS Academic

The enhanced PTE Academic test was thoroughly field tested in early 2024, and the updated test design and analyses included measures to maintain the alignment of the updated PTE Academic test to the CEFR, using both standard setting and statistical linking procedures.

The final step in the research, design and development of the enhanced PTE Academic test was to conduct a concordance study between the enhanced PTE Academic test and the IELTS Academic test.

Concordance studies between high-stakes English language tests are essential in order to support the interpretation and standards of test scores from different tests that are used for common purposes. These high-stakes tests scores carry significant currency, and it is therefore essential that test takers and test score users understand the direct relationship between test scores. The PTE Academic test and the IELTS Academic test are both used for academic, professional, and economic migration around the world. An initial concordance between the PTE Academic and IELTS Academic tests was established by Zheng & De Jong (2011). This was followed up by a revised concordance study in 2020 (Clesham & Hughes, 2020).

The primary purpose of the 2024 concordance study was to provide updated score concordance tables between the enhanced PTE Academic test and the IELTS Academic test. All the methodological steps of this concordance study followed the good practice principles developed by Knoch & Fan (2024), ensuring robust score comparisons on all four skill scores as well as the overall test score.

The full concordance study, with updated data will be published in April 2025 and includes an extended construct comparison between the enhanced PTE Academic and IELTS Academic tests, as well as overall score and communicative skill score concordance tables.

Part 1. Construct Comparison

The first component of a concordance study is to establish where the two tests have construct comparability and therefore can be linked by test score. Prior research had established the appropriacy of linking PTE Academic and IELTS Academic (Zheng & De Jong, 2011). In the light of the most recent changes, however, a new external study was commissioned to compare the construct of the enhanced PTE Academic and IELTS Academic tests.

The construct comparison was undertaken by independent researchers at the University of Bedfordshire Centre for Research in English Language Learning and Assessment (CRELLA). An excerpt from the executive summary of the research report follows:

Respond to a Situation:

The construct comparison research carried out by CRELLA presents a comprehensive analysis of the enhanced Pearson Test of English Academic (PTE Academic), examining its construct and 22 question types through the socio-cognitive framework (Weir, 2005). Test specifications, sample materials, and test-taker performances provided by Pearson were scrutinized and matched with publicly available information and previous research about IELTS Academic.

PTE Academic was found to be a rigorous test which meets or exceeds the standards expected of high-stakes academic language tests in terms of task design (context validity), cognitive processes required of test takers (cognitive validity), and the levels of test taker performance (scoring dimension). Further, its use of authentic input materials sets the test firmly in the academic domain. Its emphasis on skill integration is congruent with contemporary research which illustrates the interconnected nature of language skills.

Comparing PTE Academic and IELTS Academic, both tests are intended for higher education contexts and cover CEFR A (Basic User), B (Independent User), and C (Proficient User). For Reading, PTE Academic emphasizes diversity in passage range, whereas IELTS Academic offers greater variety in question types. Both engage major cognitive processes involved in reading, including word recognition, lexical access, syntactic parsing, and inferencing, but neither assesses intertextual reading across multiple texts. Both cover a range of reading types, providing comprehensive evaluations of reading skills. PTE Academic has a temporally shorter reading section than IELTS due to its sampling of more but shorter question types, including the often under-represented academic expeditious reading. The reliability and validity of the shorter section should be confirmed with further field and concordance testing.

Though both tests assess similar listening processes, PTE Academic focuses exclusively on academic and professional listening contexts, while IELTS Listening—overlapping with IELTS General Training—includes the personal domain relevant to more general conversations. This distinction reflects differences in priorities between academic-specific assessment and broader language use domains.

Comparison of Speaking and Writing between IELTS Academic and PTE Academic is not straightforward due to difference in the independent and integrated approaches of their test design. PTE Academic's integrated design combines Speaking and Writing into a single section—IELTS treats these skills as distinct. PTE Academic Speaking tasks require test-takers to synthesize spoken and written inputs (e.g., Describe Image; Repeat Sentence) which evaluate skills such as pronunciation, fluency, and the ability to process and convey information.

Integrated tasks (e.g., Retell Lecture; Answer Short Questions) mimic real world academic tasks and interactions, motivating the longer test time allocated to Speaking than in IELTS. For Writing, by incorporating both short and extended responses, PTE Academic engages a wide range of abilities, including concise academic summaries to detailed essay tasks. Writing tasks emphasize formal register, accurate grammar, and clarity, reflecting real world academic requirements.

Given the substantial overlap between PTE Academic and IELTS Academic, including the academic English domains targeted, the cognitive processes elicited and the level of test taker performance expected, the two tests were found to have overall comparable constructs.

Excerpt from the executive summary of the research report.

> PTE Academic was found to be a rigorous test which meets or exceeds the standards expected of high-stakes academic language tests in terms of task design (context validity), cognitive processes required of test takers (cognitive validity), and the levels of test taker performance (scoring dimension).

Part 2. Score Concordance

This part of the study involved 1,522 participants who took both the enhanced PTE Academic and IELTS Academic tests within 90 days and in a reasonably counterbalanced testing order. Participants were recruited from India, China, Australia, and the United Kingdom, representing 54 nationalities, 43 languages, and a wide range of language proficiency levels.

The tests were taken under secure test conditions to ensure accurate results. Participants were given preparation materials for both tests to ensure they were familiar with them. Analysis showed that the sample of participants was representative of the global population of PTE Academic test takers.

The concordance study followed established good-practice principles for concordance studies, ensuring robust score comparisons on all four skill scores as well as the overall test scores. The concordance tables were produced through rigorous analysis using established methodologies.

Full methodological details, statistical information and the overall and skill concordance scores between the enhanced PTE Academic Test and IELTS academic are provided in the full report.

Overall, the concordance study was more thorough than previous studies, with a larger sample size and more detailed reporting. The results provide updated and reliable score comparisons for the enhanced PTE Academic and IELTS Academic tests, helping institutions and governments make informed decisions about language proficiency.

> Comparison of Speaking and Writing between IELTS Academic and PTE Academic is not straightforward due to difference in the independent and integrated approaches of their test design.

Conclusion

This paper has described in detail how the enhanced PTE Academic test differs from the current version and outlines how the fundamental structure of the test remains the same. The updates to the test have been thoroughly supported by reference to the enhancements of the CEFR over time, as well as to the research and rationale behind the new question types and scoring rubrics.

In line with Pearson's commitment to continuous improvement, this project represents an ongoing research focus to develop new question types and scoring rubrics that complement and expand linguistic competencies in line with CEFR recommendations and enhancements, maximize the construct and face validity of the test, and encourage more spontaneous speaking and engagement with novel stimuli.

At the same time, additional research has also been carried out into the scoring of test tasks, combining the best of computer scoring with human judgment. The enhanced PTE Academic Test ensures that every speaking and writing extended response is evaluated using the most appropriate and robust scoring methodologies, combining the best of human and automated scoring.

Our research into the maintenance of test integrity and increased human oversight has resulted in the development of automated monitoring systems that flag gamed speaking and writing responses which are then considered by human expert raters.

This utilizes both the power and reach of technology, paired with the essential element of human judgment.

Therefore, although Pearson is at the forefront in terms of utilizing new technologies to score and to maintain test integrity, we also want to ensure that active human expert judgment is at the heart of our systems, combining the best of what automated systems can offer with the reassurance of nuanced human judgment.

The enhanced PTE Academic test was thoroughly field tested, producing valid and reliable outcomes. The alignment of the revised test to the CEFR was established and validated by a combination of expert standard setting judgments and statistical linking procedures.

The concordance research study established the two key elements of concordance comparability. The first element was an independent construct analysis of the enhanced PTE Academic test, carried out by the university of Bedfordshire (CRELLA). Their analyses concluded that the test was rigorous and meets or exceeds the standards expected of high-stakes academic language tests in terms of task design (context validity), cognitive processes required of test takers (cognitive validity), and the levels of test taker performance (scoring dimension). Its emphasis on skill integration is congruent with contemporary research, which illustrates the interconnected nature of language skills.

The second study established the score concordance between the enhanced PTE Academic Test and the IELTS Academic Test. The study involved 1,522 participants who took both the enhanced PTE Academic and IELTS Academic tests within 90 days. The tests were taken under secure testing conditions to ensure accurate results and robust score comparisons on all four skill scores as well as the overall test scores.

All the steps described in this report illustrate that the enhanced PTE Academic has been built on robust research foundations, provides more opportunities for test takers to demonstrate a range of essential language skills and competencies, and remains a fast, fair and reliable measure of language proficiency.

References

Bachman L., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

British Council, UKALTA, EALTA and ALTE (2022) *Aligning Language Education with the CEFR: A Handbook* https://www.ealta.eu.org/documents/resources/CEFR%20alignment%20handbook.pdf

Clesham, R., Miller, L., & Hughes, S. (2023). *Examining the relevance of PTE Academic for Australian professional bodies.* Pearson. https://www.pearsonpte.com/ctfassets/yqwtwibiobs4/6gPjOLrZorZFDeyBBUMhLt/4b6f02306bab6d303b0d83c86c178ba5/pearson-pte-australian-trg-report-june-2024.pdf

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

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching assessment (CEFR).* A manual. Language Policy Division.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment* – Companion volume, Council of Europe Publishing, Strasbourg.

Hudson, T., Detmer, E., & Brown, J.D. (1995). *Developing prototypic measures of cross-cultural pragmatics*. Second Language Teaching & Curriculum Center, University of Hawaii at Mānoa.

Hughes, S., & Clesham, R. (2024). *Test integrity Ensuring trust by integrating human judgment and AI systems.* Pearson.

Knoch, U., & Fan, J. (2024) *Test score comparison tables: How well are they serving test users?* Language Testing, 41(3) 681-693. http://doi.org//10.1177/02655322241239348

Papageorgiou, S., Stevens, R., & Goodwin, G (2012) *The Relative Difficulty of Dialogic and Monologic Input in a Second-Language Listening Comprehension Test,* Language Assessment Quarterly, 9:4, 375-397.

Roever, C. & Ikeda, N. (2022) *What scores from monologic speaking tests can(not) tell us about interactional competence.* Language Testing 39(1).

Zheng, & De Jong. (2011). *Establishing Construct and Concurrent Validity of Pearson Test of English Academic* [Research Note]. Pearson. http://pearsonpte.com/wp-content/uploads/2014/07/RN

About the authors

Dr. Rose Clesham

Dr. Rose Clesham is the Director of Assessment Research and Validity. She holds a Master's Degree in Formative and Summative Assessment from Cambridge University and a Doctorate in Educational Assessment. She is a Fellow of the Association for Educational Assessment-Europe (AEA-E) and an Honorary Associate Professor at the University of London (UCL).

She has worked extensively on OECD PISA assessments, including co-writing assessment Frameworks. Rose lectures on international educational standards, validity and reliability issues. Her research interests include the development of e-assessment and Artificial Intelligence systems, and on-going international educational strategies and policy.

Sarah Hughes

Sarah is Head of Test Design and Validation in Global Assessment at Pearson. Working in the Global Product English Language Assessment team, with a focus on research into Pearson Test of English - Academic (PTE-A) and large-scale high-stakes language testing.

Sarah has a Master's in English from Brown University. Alongside her work at Pearson she is a PhD candidate in the Cambridge Faculty of Education, researching validity and AI in assessment.

Pearson

Secure. Accurate. Trusted.

pearson.com/languages