Pearson

PTE Academic

Assessment Efficacy Report

March 2019

Contents

Introduction

Pearson's mission is to help people make progress in their lives through learning. But helping people to achieve the learning outcomes that matter most in life, like new knowledge and skills to support progression into further or higher education, or in a career, isn't something that happens by accident. It happens by design.

When we first start our learning journey, the choices of parents and educators often drive decisions about learning. As we grow, we take over from parents and educators, becoming the designers of our own lifelong learning journey. We identify the outcomes we want to achieve the most, and select learning experiences that suit how we want to access and engage with learning.

At Pearson, we are committed to supporting you to achieve the outcomes that matter most to you. That's why we design products focused on supporting the achievement of those outcomes, why we underpin the design and implementation support with evidence about what works to improve teaching and learning, and why we measure the impact of use of our products on outcomes. We use what we learn to continuously improve how our products and services are designed and used.

The 2019 Product Efficacy Reports include three audited, standards-based efficacy research studies on: Revel for Psychology, 1st edition by Marin and Hock in North America, MyPedia in India and Sistema COC in Brazil. We are simultaneously publishing non-audited efficacy reports on two of our most frequently used assessment and qualifications products — Pearson Test of English Academic and the UK regulated GCSE Maths Qualification.

We remain committed to continuously improving how we are applying efficacy in education, all with a focus on helping more people make progress in their lives through learning. This sense of purpose gives us a reason to keep on fighting, nothing spared, to improve how we do things in education.

Kate Edwards
Senior Vice President, Efficacy and Learning Research, Pearson
March 20, 2019

Special thanks
We want to thank all the customers, test-takers, research institutions and organizations we have collaborated with to date. If you are interested in partnering with us on future efficacy research, have feedback or suggestions for how we can improve, or want to discuss your approach to using or researching our assessments, we would love to hear from you at efficacy@pearson.com.

About efficacy reporting at Pearson

To be as open and transparent as possible about how we design, develop, and evaluate the impact of use of our products on learning we produce efficacy reports. We have two types of report: audited and non-audited reports.

About audited product efficacy reports

To find out more about our audited reports, go to pearson.com.

About qualifications efficacy reports

One particular type of service we have are assessments. We support our customers by designing, building, administering, scoring, and reporting on test-takers' performance in many different contexts (from K–12 classrooms to the workplace) and for different purposes (such as supporting classroom instruction, ongoing progress monitoring, or certifying fitness for employment).

Taking a test is not a learning experience in and of itself, but people who take tests are still learners on a journey. Instructors and others can use the scores and diagnostic information from assessments to make decisions about a learner's progress along their journey.

Therefore, the measure of an assessment's efficacy is not whether taking the test leads directly to higher achievement or passing the course, but whether the scores and other diagnostic information provide an accurate snapshot of what the learner knows and can do. In other words, the efficacy of an assessment is its fitness for a given purpose.

The Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014) define three attributes we can use to judge the efficacy of an assessment: validity, reliability, and fairness.

- **Validity** is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p11). Validity requires evidence that test scores can be interpreted as they are intended and can be appropriately used for a specific, defined purpose.
- **Reliability** is "the consistency of scores across replications of a testing procedure" (p33). Reliability requires evidence of the consistency of scores over time, across multiple forms of the assessment, and/or over multiple scorers.
- **Fairness** suggests that "scores have the same meaning for all individuals in the intended population" (p50). Fairness requires evidence that when assessments are administered as intended, items are not systematically biased against any particular group of test-takers and students are not hindered in demonstrating their skills by irrelevant barriers in the test administration procedures.

Given the longstanding role of these standards as a source of guidance on best practices in the development and evaluation of tests, and the role they play in the legal defensibility of assessment, Pearson has adopted these three attributes as the Assessment Quality Indicators on which we publicly report evidence underlying our assessment products.

Pearson's assessment products are designed, built, and maintained over time by teams of subject matter experts and Ph.D. level research scientists trained in the science of assessment. These teams regularly (in some cases, annually) carry out studies to collect evidence of validity, reliability, and fairness, in accordance with the Standards for Educational and Psychological Testing. This evidence is typically consolidated and published in a technical manual or technical report that is updated with each new revision of the test. We refer any interested readers to the technical manuals for full details of the research studies and associated evidence.

Product summary

Pearson Test of English Academic (PTE Academic) is a computer-based international English language test. Pearson developed PTE Academic in response to demand from higher education, governments and other customers for a test that could more accurately measure the English communication skills of international students in an academic environment. PTE Academic is accepted for study applications by thousands of academic programs around the world. It is also approved for all Australian and New Zealand student visa and migration applications.

The purpose of PTE Academic is to measure English language proficiency in the skills of listening, reading, speaking and writing. To be able to claim that PTE Academic is fit for this purpose, a variety of types of validity evidence has been collected from the various stages of test development through to its administration. 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 courses and to actively participate in the targeted tertiary level education environment.

The PTE Academic Score Report includes a candidate's overall score, communicative skills scores and enabling skills scores. The overall score reflects overall English language ability. The score is based on performance on all items in the test. The range for the overall score is 10-90 points. Scores for communicative skills (listening, reading, speaking and writing) are based on all test items that assess these skills, either as a single skill or together with other skills. The range for each communicative skill score is 10-90 points. Scores for enabling skills (grammar, oral fluency, pronunciation, spelling, vocabulary and written discourse) are based on all test items assessing one or more of these skills. The range for each enabling skill score is 10-90 points.

For reasons of transparency, it is useful to relate numerical test scores to a descriptive system that facilitates interpreting test scores in terms of predicted potential for behaviour of test takers. PTE Academic is scored in relation to the Global Scale of English (GSE), which has been aligned to the Common European Framework of Reference for Languages (CEFR). Since its publication, the CEFR has gained currency in Europe and beyond as a standard for defining, comparing and equating levels of language competence. The PTE Academic test is targeted to intermediate to advanced English language learners and assesses the full range of proficiency expressed by CEFR A1 to C2 through the GSE score range of 10-90.

The GSE is the first truly global English language standard, allowing teachers to more accurately and easily measure learner progress. Based on research involving over 6,000 teachers from more than 50 countries, the GSE extends the CEFR by pinpointing on a scale from 10 to 90 what needs to be mastered for the four skills of speaking, listening, reading and writing within a CEFR level, using a more granular approach.

PTE Academic is a language competency measure known across the world, with test centers in more than 50 countries.

Assessment quality indicators

We define efficacy in assessment using three primary assessment quality criteria – validity, reliability and fairness – as these apply to the main purpose of the assessment. The purpose of PTE Academic is to measure test-takers' academic English language competency in listening, reading, speaking and writing.

The three assessment quality criteria discussed here are the extent to which the assessment allows test users to make sound interpretations of test-takers' English language competency (validity), the consistency and accuracy of scores (reliability), and fairness of the assessments (AERA, APA and NCME, 2014).

Assessment quality indicator 1

Test scores can be interpreted as measures of English language competency and can be used for placement into academic programs or skilled migrant routes (validity).

A key PTE Academic goal is to enable test users to make sound interpretations about test-takers' English language competency. This supports identification or placement decisions by providing measures that accurately capture ability, as well as profiles of relative strengths and weaknesses across the four communicative skills.

Assessment quality indicator 2

Test scores are consistent over time and/or over multiple test administrations (reliability).

Another important goal of the PTE Academic is to minimize errors in judgment and decision-making by providing scores that are consistent over different testing occasions.

Assessment quality indicator 3

Test scores can be interpreted in the same way for test-takers of different subgroups (fairness).

PTE Academic also strives to provide scores that can be interpreted in the same way for all test-takers, regardless of gender, race/ethnicity or first language. Fairness implies that when the assessments are administered as intended, items are not systematically biased against any particular group of test-takers and test-takers are not hindered in demonstrating their skills by irrelevant barriers in the test administration procedures.

Product research

The PTE Academic team carried out studies to collect the kinds of validity, reliability, and fairness evidence described above. This evidence has been consolidated and published in a technical manual. For that reason, much of the research we summarize in the following section has been completed internally. We encourage test users who are interested in the full details of our internal research studies and associated evidence to consult the official technical manual, which is available to qualified users with appropriate credentials.

More information about the full scope of internal and external research related to PTE Academic can be found on the PTE Academic research website.

Overview of product research

Aligning PTE Academic test scores to the Common European Framework of Reference

Study citation	Pearson (2010). *Aligning PTE Academic test scores to the Common European Framework of Reference for Languages.* https://pearsonpte.com/wp-content/uploads/2014/07/Aligning_PTEA_Scores_CEF.pdf
Type of study	Concordance/alignment study
Sample size	4,028 candidates from field tests
Description of sample	Candidate performance data was used from 4,028 candidates included in the field-testing on 94 items from three item types. 147 raters were used to judge the performances against the CEFR scale.
Assessment quality indicator measured	(AQI 1) Test scores can be interpreted as measures of English language competency and can be used for placement into academic programs or skilled migrant routes (validity).

This study reports on the theoretical justification and statistical procedures used for relating PTE Academic scores to the levels of the CEFR scale, which adheres to the guidelines provided in the *Manual for relating language examinations to the Common European Framework of Reference for Languages* (Council of Europe, 2009). The study involved both a test-taker-centered approach, in which the quality of candidate responses was analyzed, and an item-centered approach, in which the intended difficulty of different items was analyzed.

For the test-taker-centered approach Pearson used test-taker responses to three item types: writing an essay, oral description of an image and oral summary of a lecture. These responses were rated on the CEFR scale by two human raters. When the raters disagreed by more than one level, a third rater was used, and the two closest ratings were kept. This data was compared to the candidate ability estimates based on the same responses scored against PTE Academic criteria. The results showed similar increases in median PTE Academic ability and CEFR level, with overlap between the levels. This overlap is expected, given the way language learning ability is expressed by the CEFR and the original scaling of the CEFR levels (North, 2000). Learners of a given level of ability are estimated to be most likely at a certain level, but this does not mean their probability of being at an adjacent level is zero.

From this data, cut-off points could be determined for PTE Academic scores in relation to the CEFR levels. The original scaling of the levels (North, 2000) was based on the Rasch model and cut-offs were defined at 0.5 probability of being 'at a level'. The computer program FACETS (Linacre 1995,2005) was used to analyze the probabilities of different abilities being 'at a level' along each scale. The correlation between the two measures was 0.69, with a polynomial regression found to be better fitting than a linear regression. Following this, the CEFR lower bounds were expressed on the PTE Academic ability scale using an equipercentile equation, which yielded a concordance table that demonstrates the relationship of candidate abilities expressed by the two scales.

For the item-centered approach, item writers indicated intended CEFR targeting for each item during development. Mean scores were calculated for each of the intended levels. These intended mean difficulties can be related to the CEFR level calculated in the test-taker-centered approach. Both estimates were derived independently and show high levels of agreement ($r = 0.99$).

This study also includes a discussion about the meaning of 'being at a level' on the CEFR, as different interpretations would impact the way in which CEFR levels and PTE Academic scores can be related mathematically. Pearson defines 'being at a level' as 'the ability threshold at which it is more likely than not for a person to be successful in performing any task at that level'. This study demonstrates how these threshold abilities can be equated between the CEFR and PTE Academic scales and further supports the validity of the PTE Academic test as a measurement of language ability that is interpretable for its intended use.

Alignment of the Global Scale of English to other scales: the concordance between PTE Academic, IELTS and TOEFL

Study citation	De Jong, J. and Benigno, V. (2017). *Alignment of the Global Scale of English to other scales: the concordance between PTE Academic, IELTS and TOEFL*.
Type of study	Concordance/alignment study
Sample size	3,197 test-takers
Description of sample	The sample was drawn from the field-testing program, in which candidates were requested to provide self-reported scores and official score reports from other language tests they had taken within two months of participating in the PTE Academic field test. Approximately 1 in 4 candidates who self-reported scores also submitted an official score report.
Assessment quality indicator measured	(AQI 1) Test scores can be interpreted as measures of English language competency and can be used for placement into academic programs or skilled migrant routes (validity).

This paper reports on the methodology used to establish scale concordance between PTE Academic's Global Scale of English (GSE), the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS). Both IELTS and TOEFL are established internationally recognized English language assessments used in academic selection and migration processes. Both assessments report sub-scores for reading, writing, listening and speaking skills, as well as an overall score, just as PTE Academic. Each assessment uses its own scale: IELTS (0-9) and TOEFL (0-120 points).

During the field-testing program for PTE Academic, candidates were asked to self-report the scores they had received on other language tests taken within two months of participating in the field test, and to provide official score reports. Approximately 1 in 4 candidates submitted official score reports. However, evidence supports the validity of the self-reported scores. The correlation between the self-reported results and the official score reports was .82 for TOEFL iBT and .89 for IELTS.

In the first phase of field-testing, candidates self-reported 327 scores for TOEIC, 339 scores for TOEFL (including PBT, CBT and iBT versions), and 2432 scores for IELTS. The correlation between these scores and the candidates' PTE Academic scores ranged from 0.46 (TOEFL CBT) to 0.76 (TOEIC and IELTS). From this set of data, concordance coefficients were produced by linear regression and used to

predict candidates' scores on IELTS and TOEFL iBT, based on their PTE Academic scores.

In the beta testing that followed the field-testing, an additional 42 TOEFL iBT scores and 57 IELTS scores were collected from test-takers. The correlations between these reported scores and scores predicted based on PTE Academic scores were 0.73 (IELTS self-reported), 0.75 (TOEFL iBT self-reported), 0.77 (TOEFL iBT official report) and 0.83 (IELTS official report).

This results from this study formed the preliminary basis for the scale concordance between the GSE, IELTS, and TOEFL, and has been further validated by independent research aligning TOEFL iBT to IELTS and the CEFR. This concordance study supports the validity of PTE Academic as a measurement of language ability, particularly in the context of its comparability to other measures of language ability used in university admissions selection processes.

Standard setting study – concordance with the Canadian Language Benchmarks (CLB)

Study citation	Jones, G., De Jong, J., Zheng, Y., Booth, D., Strachan, A (Rev.) (2017). *A Standard Setting Study to Establish Concordance between the Pearson Test of English Academic (PTE A) and the Canadian Language Benchmarks (CLB).* https://pearsonpte.com/wp-content/uploads/2018/07/Pearson-Standard-Setting-Study-between-PTEA-and-CLB.pdf
Type of study	Concordance/alignment study (modified Angoff and Contrasting Groups methods)
Sample size	23 experienced English language practitioners
Description of sample	The panelists were drawn from different regions of Canada to represent geographical diversity, with the majority coming from Ontario, where the standard setting workshop took place. Panelists represented the range of educational establishments where the CLB are used, and the vast majority of the panelists had experience of using the CLB for five years or more. The key criterion for selection was familiarity with the CLB.
Assessment quality indicator measured	(AQI 1) Test scores can be interpreted as measures of English language competency and can be used for placement into academic programs or skilled migrant routes (validity).

The aim of this study was to establish a link between PTE Academic scores and the Canadian Language Benchmarks (CLB) in order to determine a set of cut scores that correspond to the respective boundaries between the CLB levels 4 to 10 for listening, reading, speaking and writing. The CLB are recognized as the official Canadian standards of English language proficiency for immigration purposes. The CLB were developed by the Centre for Canadian Language Benchmarks and are owned by the IRCC (Immigration, Refugees and Citizenship Canada). The standard-setting methodology used in this study has been similarly used to set cut scores for work or migration purposes in other language proficiency contexts.

PTE Academic items were selected to represent the full range of difficulty, which had been established through robust field-testing and live test administration.

For listening and reading (receptive skills), panelists were asked to independently and anonymously review a selection of the PTE Academic items and determine which CLB best described a learner who had a 50% chance of answering the item correctly. For speaking and writing (productive skills), panelists were asked to independently and anonymously review responses to a selection of PTE Academic items and determine which CLB best described the English language proficiency of the speaker/writer.

Four regression functions were computed, one for each of the skills, to relate the CLB ratings from the workshop to the PTE Academic Global Scales of English (GSE) values previously established for the items. Using the regression functions, the GSE value for each CLB was computed for the skills scores and overall score.

The resulting scale was validated in relation to previous concordance studies linking the CLB to IELTS and IELTS to PTE Academic. The relationship that can be inferred from previous concordance studies was supported by the findings from the standard-setting workshop, with an error of measurement less than or equal to 3 points on the GSE scale.

The outcomes of the workshop provide a robust basis for establishing the cut-off points on the PTE Academic scale for different levels on the CLB.

Automated scoring whitepaper

Study citation	Pearson (2019). *Pearson Test of English Academic: Automated Scoring. Automated Scoring Whitepaper.* https://pearsonpte.com/wp-content/uploads/2018/06/Pearson-Test-of-English-Academic-Automated-Scoring-White-Paper-May-2018.pdf
Type of study	Reliability analysis
Sample size	>10,000 test-takers >50,000 written responses >400,000 spoken responses >200 human raters
Description of sample	The written and spoken responses were collected during PTE Academic field-testing. The raters were drawn from Australia, the United Kingdom and the United States.
Assessment quality indicator measured	(AQI 1) Test scores can be interpreted as measures of English language competency and can be used for placement into academic programs or skilled migrant routes (validity). (AQI 2) Test scores are consistent over time and/or over multiple test administrations (reliability). (AQI 3) Test scores can be interpreted the same way for test-takers of different subgroups (fairness).

This whitepaper describes the underlying technologies used to power the automated scoring of spoken and written responses and explains the ways in which these technologies have been developed and tested for PTE Academic.

Written responses are scored using Intelligent Essay Assessor (IEA), which is powered by the Knowledge Analysis Technologies (KAT) engine. Test-takers produce written responses that range from approximately 50 to 300 words. During field-testing, more than 50,000 of these responses were each scored by two randomly selected human raters along traits related to content, formal requirements, grammar, vocabulary, general linguistic range, spelling, development, structure and coherence. This large set of data was used as input to train the IEA to score test-taker responses at trait level. The overall writing score for each test-taker is a sum of the trait scores for all of the writing items. For this overall writing score, the correlation between the human score and the

machine-generated score was 0.88, which is slightly higher than the correlation between pairs of human raters, 0.87.

Spoken responses are scored using the Ordinate technology. During field-testing, more than 400,000 spoken responses were scored by human raters along traits related to content, vocabulary, language use, pronunciation, fluency and intonation. The responses were produced by a diverse population of over 10,000 test-takers, including 126 different accents. These scores were used as input to the advanced speech processing system, which then built scoring models to relate all of the scoring engine's observations to the human scores. The correlation between the human scores and the machine scores for the overall measure of speaking was 0.96

The automated scoring systems for PTE Academic show high correlations with scores generated by expert human raters. This means that the automated system acts like a human rater when assessing test-takers' language skills, but does so with the precision, consistency and objectivity of a machine. By automating the scoring process, PTE Academic is able to ensure that test-takers' scores are derived in exactly the same way for every candidate. This level of analysis has been incorporated into the development for all new PTE Academic speaking and writing items, ensuring that the automated scoring reflects human judgements of language ability and is applied consistently and fairly for all test-takers.

Differential item functioning and unidimensionality

Study citation	Pae, H. (2011). *Differential item functioning and unidimensionality in the Pearson Test of English Academic. Pearson Research Note. https://pearsonpte.com/wp-content/uploads/2014/07/RN_Differential-ItemFunctioning.pdf*
Type of study	Differential item and test functioning; dimensionality analysis
Sample size	140 English language learners (ELLs)
Description of sample	The participants included 140 ELLs who participated in Field Test I, with a mean age of 26.45 (SD=5.82). The gender balance was 53.6% female and 46.5% male. The proportion of test-takers who had lived in English-speaking countries was 64%.
Assessment quality indicator measured	(AQI 1) Test scores can be interpreted as measures of English language competency and can be used for placement into academic programs or skilled migrant routes (validity). (AQI 3) Test scores can be interpreted the same way for test-takers of different subgroups (fairness).

This study aimed to answer two research questions related to the construct validity and fairness of the PTE Academic test.

1. Do the items in PTE Academic behave the same for gender subgroups (male and female) and for English Language Learners (ELLs) from different language learning contexts (English speaking countries and non-English speaking countries)?
2. To what extent do the item responses of PTE Academic form a unidimensional construct, meaning that they measure a unitary latent trait underlying language ability?

Data from one test form used in Field Test I was used to explore these research questions. The selected test form included 86 items.

To address the first research question, the psychometric properties of the items were analyzed using a Partial Credit Model (Masters, 1982) in the Winsteps software (Linacre, 2010). differential test functioning (DTF) and differential item functioning (DIF) analysis techniques were used to determine if the test functioned differently for different subgroups. For gender, the difficulty of each item for both males and females was remarkably similar. Of the 86 test items, only two were identified as outliers, and item difficulties for both males and females had

correlations of 0.84 (0.98 measurement-error-removed disattenuated correlation). For English-speaking and non-English speaking learning contexts, there was similarly no evidence of test bias. Of the 86 items, only one was identified as an outlier. While outliers were identified, we cannot assume to know the cause of the outlying position, and the overwhelming majority of items performed identically between gender and learning context subgroups. PTE Academic items also underwent a separate sensitivity review to ensure the item content does not induce bias (see p. 12 Item Sensitivity Review for qualitative analysis of item content).

To address the second research question, the data was analyzed using principal component analysis (PCA) to determine if the test responses reflected a unidimensional construct. The amount of variance explained by components in the data was 76.1% (20.7% persons and 55.4% items), which exceeds the minimum requirement to demonstrate the presence of a dominant first factor. These findings suggest that there were no meaningful components beyond the primary dimension of measurement. These findings are in line with the dimensionality analyses carried out during field-testing (see pp. 13-15) and in a separate external study (Reckase, M. & Jing-Ru, X. (2014), which conclude that, while some dimensionality is present in the data, a unidimensional model fits the scoring data well. This supports the validity of separate score reporting on the basis of substantive content considerations rather than statistical considerations. Further research continues in this area.

Overall, these findings demonstrate that the PTE Academic test, as a whole, produces ability estimates with no meaningful differences between different genders and learning contexts, suggesting that PTE Academic scores can be interpreted in the same way for members of these subgroups. The test also appears to measure a unitary latent trait. Taken together, these findings indicate that PTE Academic measures a unidimensional construct fairly across subgroups, which further supports the test's argument for construct validity.

Item sensitivity review

Study citation	Pearson (2018). *Pearson Test of English Academic Technical Manual (pp. 50-54)*
Type of study	Sensitivity review and DIF analysis
Sample size	16 reviewers
Description of sample	The reviewing panel comprised one chair and 15 reviewers, who were all highly proficient in English and had experience in teaching English as an additional language and, in some cases, test development. The panel represented different nationalities and regions.
Assessment quality indicator measured	(AQI 1) Test scores can be interpreted as measures of English language competency and can be used for placement into academic programs or skilled migrant routes (validity). (AQI 3) Test scores can be interpreted the same way for test-takers of different subgroups (fairness).

A sensitivity review was conducted on PTE Academic test items throughout item development and field-testing. The purpose of the review was to detect and remedy any potential instances of bias or construct-irrelevant variance stemming from content that could elicit a strong emotional response for groups of test-takers. The review comprised three phases and sought to identify sensitive content related to:

- cultures
- religions
- ethnic groups
- socio-economic groups
- people with disabilities
- gender roles
- use of positive language
- field-specific knowledge

In the first phase, the 16 panelists were instructed to review items and the accompanying stimulus material to determine if there was no sensitive content, if sensitive content was present but could be edited, or if sensitive content was present and could not be edited. Each item was reviewed by two panelists and 83.5% of the items were determined to have no sensitivity issues by both panelists. The chair reviewed the remaining items, judging that only about 2% of

the items should be removed from the bank, as they could not be edited to remedy the sensitivity issues.

The second phase focused on the remaining 14.5% of items from Phase I. This phase included an analysis of the test-taker demographics in relation to performance on Field Test I. There was sufficient data on test-takers' gender, region of birth, and field of study to run a differential item functioning (DIF) analysis for a subset of the items. Some items showed a statistically significant difference between groups and were removed from the item bank, totaling approximately 0.4% of the item bank.

The third phase included a qualitative internal analysis for the subset of items that were not amendable to DIF analysis in Phase II. Two internal experts reviewed the items independently to determine if they should be kept, edited or removed from the item bank. Of these, 63% were kept or edited and the remainder were removed from the bank.

The sensitivity review ensured that items used in further field-testing would not introduce construct-irrelevant variance or systematic bias by inducing a strong emotional response in groups of test-takers. This process also helped form the review guidelines that are employed in the ongoing item development for PTE Academic, which supports the continued validity and fairness of the assessment.

Field Test I

Study citation	Pearson (2018). *Pearson Test of English Academic Technical Manual (pp. 81-145)*
Type of study	Field test
Sample size	6,208 test-takers
Description of sample	The sample test-taker population was wide-ranging and diverse. Test-takers reported 126 different countries of citizenship and 91 different primary languages. Approximately two-thirds of test-takers reported having a bachelor's degree or higher level of education and most test-takers were aged 21 or older. Recruiting for the field test took place primarily among the international student population of Australian universities.
Assessment quality indicator measured	(AQI 1) Test scores can be interpreted as measures of English language competency and can be used for placement into academic programs or skilled migrant routes (validity). (AQI 2) Test scores are consistent over time and/or over multiple test administrations (reliability). (AQI 3) Test scores can be interpreted the same way for test-takers different subgroups (fairness).

The main test development phase for PTE Academic involved the design and implementation of two large scale field tests and a beta test. Field-testing was used to iteratively refine assessment functioning and the beta test was implemented to ensure delivery system functionality. Field Test I had seven objectives:

1. Collect data for training and validating automated scoring systems
2. Determine scoring model for each item type
3. Eliminate items that have substandard quality
4. Establish method for aggregating item scores and generating score reports
5. Establish minimum number of items and item types to yield reliable and valid measurement
6. Estimate item difficulty parameters to build item bank
7. Collect data on test-taker demographics, opinions of the test, and testing behaviors

Field-testing was conducted in participating Pearson VUE testing centers internationally. One of 38 test forms was assigned randomly to a test-taker during the registration process. The test forms contained overlapping subsets of items to administer to different subsamples of test-takers, such that all items occurred in at least two test forms. To reduce the risk of losing information on items because they were not reached, the items were divided into three blocks of about equal length in estimated time to complete and the ordering of blocks varied over different test forms. Items were selected across the range of CEFR levels as specified by the item writer and development team.

The test administered 95 items within a 195-minute time limit. Test-takers were given a survey on completion of the test, which included questions on test-taker satisfaction with test instructions, test difficulty and overall impression of the test. This information was analyzed and used to improve the quality of the test.

The sample population for Field Test I included 6,208 test-takers recruited primarily from the international student population at Australian universities. The test-takers reported 126 different countries of origin and 91 different first languages. The sample was selected to represent a wide range of test-taker language backgrounds so that data could be collected to train the automated speaking system.

The assessment data collected during Field Test I was analyzed using classical analysis, factor analysis and IRT analysis to refine item scoring models, establish methods for aggregating scores, determine the optimum test structure, and estimate item difficulty parameters on which to build the item bank.

Overall, Field test I yielded valuable results that provided a solid basis for further development of the PTE Academic test. A number of revisions and adjustments were made to the test, which could then be further analyzed in Field Test II, including:

1. Refine task layout, instructions and item difficulty
2. Define more specific item design parameters and scoring rubrics
3. Adjust the number of items and duration of test
4. Improve test audio equipment and test-taking environment

The efforts of Field Test I demonstrate that the validity, reliability, and fairness of the assessment were thoroughly investigated throughout the iterative development and field-testing phases, providing substantial evidence toward the quality of the assessment.

Field Test II

Study citation	Pearson (2018). *Pearson Test of English Academic Technical Manual (pp. 145-242)*
Type of study	Field test
Sample size	4,172 test-takers
Description of sample	The sample test-taker population was wide-ranging and diverse. Test-taker demographics were similar to Field Test I. Sample overlap with Field Test I included approximately 11% of items and almost 900 test-takers.
Assessment quality indicator measured	(AQI 1) Test scores can be interpreted as measures of English language competency and can be used for placement into academic programs or skilled migrant routes (validity). (AQI 2) Test scores are consistent over time and/or over multiple test administrations (reliability). (AQI 3) Test scores can be interpreted the same way for test-takers different subgroups (fairness).

Field Test II built on the valuable information collected in Field Test I and had five objectives:

1. Evaluate the success of automated scoring and refine the process of aggregating trait scores
2. Assess item functioning via classical and IRT analysis
3. Finalize methodology for calculating enabling skill scores and scaled scores
4. Review the relationship between CEFR levels and field-testing data
5. Finalize test composition criteria – test time, item seeding, test information and standard error of measurement (SEM)

Field Test II was taken by over 4,000 candidates with similar demographics as Field Test I. The spread over the 38 test forms was well balanced and resulted in over 200 administrations per item. Overlap between Field Test I was created both by incorporating approximately 11% items per item type from Field Test I in Field Test II. In addition, a person overlap was created because close to 900 candidates who had participated in Field Test I also participated in Field Test II.

More intensive training of the human raters in Field Test II, combined with the use of a larger number of traits for scoring written and spoken responses, led to an improvement in the scores obtained from the automated scoring systems.

The assessment data from Field Test II was analyzed using both classical and IRT analysis to investigate and identify optimal scoring rules. After deciding on the scoring rules, the total data set was inspected with classical item statistics leading to the removal of a small number (1%) of candidates and about 9.4% of the items. At test-level, the IRT results corroborated the findings from Field Test I, demonstrating a high internal consistency and an item distribution well-targeted at the candidate population. Classical reliability estimates were calculated for odd/even item split halves, with overall reliability of 0.96 and skill score reliability ranging from 0.89 to 0.94.

Enabling skill scores were investigated in relation to ability estimates. The results showed general homogeneity of the skill score/ability relationship, suggesting that enabling skill scores could be reported in context. Initial investigation of the relationship of speaking and writing scores with the matching CEFR scales shows a positive relationship, as found in Field Test I (see p. 7 Aligning PTE Academic test scores to the Common European Framework of Reference for the final CEFR analysis). In general, the results from Field Test II corroborate those obtained from Field Test I and, when combined, supported final decisions about test composition.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education and Joint Committee on Standards for Educational and Psychological Testing (US) (2014). *Standards for educational and psychological testing*.

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

Council of Europe (2009). *Manual for relating Language Examinations to the Common European Framework of Reference for Languages*.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

Linacre, J. M. (2005) Facets computer program for many-facet Rasch measurement. Beaverton, Oregon.

Linacre, J.M. (2010). Winsteps® [Computer Software]. Beaverton, Oregon.

North, B. (2000). *The development of a common framework scale of language proficiency (Volume 8 of Theoretical studies in second language acquisition)*. New York: Peter Lang.

Reckase, M. D. and Xu, J.-R. (2015). The Evidence for a Subscore Structure in a Test of English Language Competency for English Language Learners. *Educational and Psychological Measurement*, 75(5), 805–825.

Pearson