

Validity and Reliability in PTE Academic

Introduction

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 will more accurately measure the English communication skills of international students in an academic environment. The purpose of this test is to measure test takers' academic English language competency in listening, reading, speaking and writing. PTE Academic is endorsed by the Graduate Management Admission Council® (GMAC®). GMAC is the owner of the Graduate Management Admission Test® (GMAT®).

To develop this comprehensive computer-based English language test, Pearson worked with internal and external test development experts. In addition, the company conducted an extensive field test program to test the items of PTE Academic's test items and measure their effectiveness in assessing a test taker's ability to communicate in English in an academic environment.

Validity is the extent to which the test results are relevant and meaningful for the purpose of the test.

Reliability is the extent to which the test results can be relied upon, i.e., that the results will be similar on repeated occasions.

Modern test design

PTE Academic is a multi-level, integrated-skills test of English language proficiency. It is designed to assess English language competence set in the context of academic programs of study that are available around the world. The development of the tasks and items for the test followed consultation with external stakeholders and test development professionals. PTE Academic is supported by two external advisory boards composed of experts in applied linguistics and assessment, who have overseen the development of the test from a professional perspective.

PTE Academic consists of 20 item types reflecting different modes of language use and setting different response tasks or response formats. The maximum duration of the test is 3 hours.

PTE Academic scores are delivered to test takers online, typically within 5 business days. There are two versions of the score report: one for the test taker and an official institution version. The score report provides three types of scores: an overall score, scores for communicative skills (listening, reading, speaking and writing) and scores for enabling skills (i.e. grammar, oral fluency, pronunciation, spelling, vocabulary and written discourse). The score scale ranges from 10 to 90.

A number of innovative features have been integrated into the design of PTE Academic. These developments build on advances in the understanding of language competence in the academic environment, including the importance of integrated skills, and advances in technology, particularly in the areas of scoring and reporting results. As the worldwide leader in publishing and assessment for education, Pearson is using several of its proven, proprietary, patented technologies to automatically score test takers' written and spoken performance on PTE Academic.

Major aspects of PTE Academic which contribute to the reliability of the measurement of the test and the validity of the test content and purpose, include:

- high levels of objectivity, including the automated marking of items;
- high numbers of scoring points (+ 100) on each of the four communicative skills;
- detailed peer review and sensitivity analysis included as part of test development procedures;
- integrated skills item types that reflect the context in which the test taker will use English;
- quality assurance procedures at every stage, including detailed analysis of item functioning as well as item and task difficulty;
- item banking that ensures equated test papers and test security;
- a wide-ranging research agenda with conference presentations and peer-reviewed publications;
- test development procedures and research open to external scrutiny.

Initial test reliability studies

PTE Academic is the focus of research to ensure the reliability of its scores. Below are two examples of studies that have been conducted on samples drawn from the large-scale, two-part field testing program that Pearson implemented to inform the development of the test.

Study 1: Reliability of human ratings and machine-generated scores

After calibrating the automated scoring engines, an independent set of test takers' data (which had not been used for training the scoring engines) was used to investigate the reliability of human ratings and machine-generated scores, and the degree to which the machine generated scores can predict human ratings. The overall reliability for both human marks and for machine marks was 0.97. This is a very high coefficient and demonstrates the accuracy of both human and machine marking. The correlation between human and machine marking was also very high, at 0.96, indicating that the machine-generated scores explain 92% of the variance of the human ratings.

Overall, this study demonstrated that the automated model can predict human ratings to a high degree for this test. In the future, randomly selected samples from live administrations of PTE Academic will be rescored to ensure that the high level of accuracy of marking is maintained.

Study 2: Reliability at test form and skill level

Reliability statistics were also collected for overall scores and the communicative skills scores on PTE Academic field tests. For field test 2, the scores on two independent halves of a test (odd and even items) was computed. The correlation between these half-test scores for the overall test scores was found to be 0.92 and around 0.82 for each of the communicative skills. Based on these correlations, the reliabilities for a full length test can be estimated using the Spearman-Brown prophecy formula at 0.96 for the overall test score and around 0.90 for each of the communicative skills.

These measures, based on field test data, already indicate a high level of reliability of the test forms for the whole test, and also for each of the communicative skills. By careful selection of items from both field tests for inclusion in the item bank from which live tests are randomly drawn, the overall test reliability for live test forms was estimated to average at 0.97 (± 0.01) for the overall score, and a range of 0.91 to 0.92 for the communicative skills.

	Overall	Reading	Writing	Listening	Speaking
r_{tt}	0.97	0.92	0.91	0.91	0.91

Reliability coefficients for overall test score and communicative skill scores in live test forms

In the area relevant for university admission decisions (scores from 43 to 75), this leads to an error of measurement that is around 2.5 score points on the overall score and less than 4.5 points on the scores for each of the communicative skills.

Ensuring test validity

To ensure that PTE Academic is valid and fit for purpose, evidence was collected from the first stages of test development and throughout implementation and launch. It was important that the test was linked to other external frameworks of language proficiency, so PTE Academic was benchmarked to the levels of language proficiency defined in the Common European Framework of Reference for Languages (CEF), from the test development stage. The CEF levels constitute a worldwide benchmark for language ability. It was developed by the Council of Europe (2001) to enable language learners, teachers, universities or potential employers to compare and relate language qualifications by level. To establish alignment with the CEF, several steps have been taken:

- Test item writers received specific training in understanding, interpreting and using the CEF. They were then asked to consider the CEF as the construct model for the test design and provide estimates of difficulty in terms of the CEF levels for each item submitted.
- CEF pre-estimates were reviewed by test experts, revised if necessary and stored in the item bank.
- DIALANG test items were included as anchors in the field tests.
- Student responses were rated according to the CEF levels by human raters, independently of the test scores.
- Additional concordance studies have been devised that will provide further concurrent validity evidence by comparing PTE Academic scores with other major English language tests, i.e., TOEFL and IELTS scores. Preliminary estimates have provided a concordance range for university admission requirements for the three tests, and the preliminary results are available at www.pearsonpte.com/PTEAcademic/Pages/TestScores.

PTE Academic targets the population of students seeking admission to international education programs where the primary language of instruction is English. These programs vary considerably in terms of content, level and structure. To support these diverse institutions, PTE Academic is based around general academic contexts, rather than the specific intended field of study of test takers. Therefore, topics were selected to cover a wide range of academic contexts but avoiding texts that require specific domain knowledge that may cause bias toward or against certain test takers.

All test items and tasks undergo a series of review processes at multiple stages, including author and peer review, sensitivity analysis, internal review, item migration quality checks and item pool review. Similarly, the psychometric properties of each task and item are ascertained before they can be included in the item bank from which test versions are drawn by random sampling from the different task types that together define the test construct.

Extensive field testing and analysis

Comprehensive procedures, including two rounds of large-scale field tests and follow-up surveys, have been carried out to ensure the quality of PTE Academic. Field test 1 (August to October 2007) had a sample size of 6,208, and field test 2 (May to June 2008) had a sample size of 4,169. These field tests were conducted with students who had a similar level of language proficiency to that of prospective PTE Academic test takers. A number of English native speakers were recruited as a control group for item statistics comparison. For the field test data to be representative of the actual test taker population, careful steps were taken to exclude test taker data with irregularities in test administrations, as well as data of test takers identified as insufficiently motivated.

The psychometric analysis of PTE Academic field test data:

- helped determine scoring models;
- provided data to be used for training and validating intelligent automated scoring systems;
- identified items with substandard quality, which were then eliminated;
- established how item scores can be combined to generate reporting scores;
- established the minimum number of items for each item type in the test structure;
- defined difficulty parameters for all items.

As part of Pearson's commitment to continuous improvement, every effort is made to enhance test design. For example, analysis of the survey results has led to a number of revisions and adjustments to PTE Academic. The revisions include optimizing the timing of the test items, adjusting testing facilities and task instructions, tailoring item difficulty to the target test taker population, tightening test specifications and refining test length, test structure and item layout.

In addition, PTE Academic is supported by a growing research agenda. Data from the field test program have been used by Pearson staff and also by external experts to support investigative studies into various aspects of the development of PTE Academic. Information about our research projects can be found at www.pearsonpte.com/research/Pages/home. Pearson will further develop this strong research base in order to further the company's understanding of the complexities inherent in the assessment of English language proficiency and to support its recognition in the academic and assessment communities.

For further information about PTE Academic visit www.pearsonpte.com or email PLTsupport@pearson.com.

Reference

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