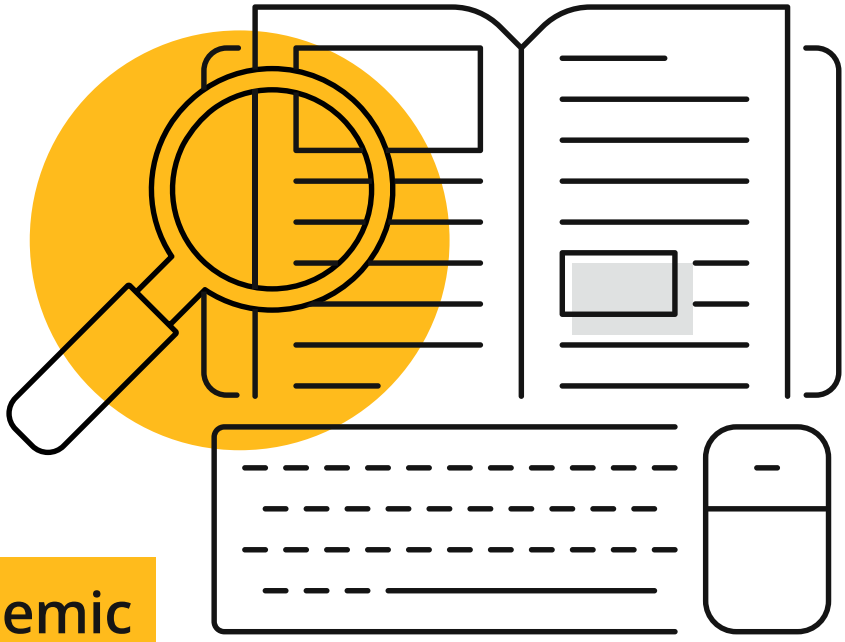
Pearson
PTE Academic



PTE Academic

Research Summary of Shortened Test Form

Dr. Rose Clesham. July 2021

Background

Pearson Test of English Academic was launched in November 2009 and represented the state of the art in language testing and operationalizing the latest insights in language testing constructs as well as implementing advanced technologies in collecting student responses and calibrating linguistic proficiency.

Although PTE Academic has maintained a substantive lead in terms of scoring and realizing the highest level of reliability in the industry, there is continual research and updating of the test, using information on item level performance.

PTE Academic uses 20 item types to sample students' language abilities and report these as an overall score, scores for the four communicative skills and scores for six enabling skills. Past research has shown that some item types contribute only marginally to the information needed to generate reliable scores, particularly when they are used multiple times over a test.

PTE Academic in its current form is designed to take a maximum of 180 minutes, which includes time taken to answer seed items. Seed items are trial items and are not used to score the test taker. There are various reasons why the reduction of this testing time has been a focus of research over the past few years.

Summative tests need to assess the domain of language proficiency with accuracy and reliability; however, over-assessment should be avoided. High stakes testing can be a stressful experience for test takers. If enough information is collected to assign test takers accurate test scores, then extended testing times do not add to scoring information. As an array of scoring data is collected on every item type across the test, research has been carried out over time to monitor how efficient items are in terms of how much measurement information they provide.

It is not simply a matter of item efficiency. PTE Academic prides itself on providing a valid and authentic experience in terms of language testing. Therefore, there are items that are open ended and have multiple opportunities to gain score points. There are also shorter, more targeted items and integrated tasks across communicative skills (e.g. listening/speaking; reading/writing) where the purpose is to assess different language skills in the context of natural language use.

This paper outlines research developed over several years that has informed the revised test model for PTE Academic.

Research evidence

As PTE Academic collects item level data, studies have been carried out over time exploring scoring patterns. Some of these studies have been carried out in-house, however a number of studies have been carried by external researchers, and in particular by Dr. Ying Zheng from Southampton University. For the purposes of these studies, data sets of 10000 test takers over 31 different test forms were used for analyses. These studies include:

1. Zheng, Ying and Mohammadi, Shaida (2013) An investigation into the writing construct(s) measured in Pearson Test of English Academic. *Dutch Journal of Applied Linguistics*, 2 (1), 108-125. (doi:10.1075/dujal.2.1.10zhe).

Pearson Test of English Academic (PTE Academic) has six item types that assess academic writing either independently or integratively. This research focused on evaluating the construct validity and effectiveness of the six writing item types. Exploratory Factor Analysis was performed to examine the underlying writing constructs as measured by the six item types. Item scores for different writing skills were subjected to Rasch IRT analysis. The difficulty of the item types was estimated and the effectiveness of each item type was evaluated by calculating the information function of each one. The results identified two writing constructs: an Analytical/Local Writing construct and a Synthetic/Global Writing construct. The study was used to explore the use of multiple item types and their effectiveness, and for test users on how they can improve their writing skills.

2. Ying Zheng (2018-19) An investigation of item type efficiency and construct relevance in the speaking and reading items of PTE Academic. PTE Research.

This two-stage study used the parameters of test construct relevance and item efficiency to examine speaking and reading item types on PTE Academic. The purpose of the study was to inform the design of the next generation of PTE Academic using statistical evidence to model item difficulty, item discrimination, as well as item efficiency and quality.

3. Ying Zheng (2020-21) PTE-Academic- next phase: Perspectives from trait marking and item type efficiency. PTE Research.

This project consisted of three stages. The first stage reviewed current language testing research to identify alternative task types which are suitable for the assessment of academic English. Focusing on assessment of spoken language in authentic contexts through single or integrated skills assessment. The second stage further elaborated on a study of single trait versus multi trait assessment of language ability. The third stage conducted a detailed analysis using recent data to evaluate item type efficiency for PTE Academic. The aim of the project was to gather literature information and empirical data from the perspectives of trait marking approaches and item type efficiency indices to inform the design of the next phase of PTE Academic.

Research Outcomes

On the basis of these research projects, various models were then produced to create a number of potential new test models. Although the research was focused on shortening overall testing time, it was an essential condition that any proposed new test model assesses with high levels of validity and reliability necessary for the purposes of this test. This means that each test model was interrogated in terms of adequate language skill coverage, retention of all item types and ensuring that reliability metrics for the four skills and the overall test remain very high. In addition, the Standard Error of Measurement (SEM) was required to remain very low.

New Test Model, excluding seed items

The table below shows the selected model. The model was selected based on its ability to reduce testing time while demonstrating high levels of reliability and maintaining the balance of assessed skills. Highlighted rows indicate item type reduction.

Item Type	Section	Skill (int)	Skill (primary)	Item Type Description	Live Model	Revised Model
07-SR-READ	A	SR	Speaking	Read Aloud	6	6
16-LS-REPT	A	LS	Speaking	Repeat Sentence	10	10
19-SS-DESC	A	SS	Speaking	Describe Image	6	3
20-LS-PRES	A	LS	Speaking	Retell Lecture	3	1
21-LS-SAQS	A	LS	Speaking	Answer short question	10	5
08-RW-SUMM	A	RW	Writing	Summarize written text	2	1
17-WW-ESSA	A	WW	Writing	Write essay	1	1
18-RW-GAPS	B	RW	Reading	Reading & Writing: fill the blanks	5	5
02-RR-MAMC	B	RR	Reading	Reading: Multiple choice, multiple answer	2	1
04-RR-DRDR	B	RR	Reading	Re-order paragraphs	2	2
05-RR-GAPS	B	RR	Reading	Reading: Fill the blanks	4	4
01-RR-SAMC	B	RR	Reading	Reading: Multiple choice, single answer	2	1
15-LW-SUMM	C	LW	Writing	Summarize spoken text	2	1
10-LL-MAMC	C	LL	Listening	Listening: Multiple choice, multiple answer	2	1
13-LW-GAPS	C	LW	Listening	Listening & Writing: fill in the blanks	2	2
06-LR-HILI	C	LR	Listening	Highlight correct summary	2	1
09-LL-SAMC	C	LL	Listening	Listening: Multiple choice, single answer	2	1
11-LL-GAPS	C	LL	Listening	Select missing word	2	1
12-LR-HOTS	C	LR	Listening	Hotspots	2	1
14-LW-DICT	C	LW	Listening	Dictation	3	3
				TOTAL Items	**70**	**51**
				TOTAL Time (min)	**139.7**	**99**
				TOTAL Time reduction (min)	**0**	**40.7**
				Section A time (min)	**67.1**	**47.8**
				Section B time (min)	**31.1**	**27.1**
				Section C time (min)	**41.5**	**24.1**

Comparison of overall testing time including seed items

	Average total test time (minutes - excl seeds)	Reduction in test time (minutes)	Average total test time (minutes - incl seeds)	item #
Current Test	139.7	/	158.4	70
New Test	99	40.7	117.7	51

The test time is reduced by approximately 40 minutes under the new model, while maintaining the same amount of seed time to ensure no disruption of item development and trialling of new items.

Comparison of internal reliability of the tests (Cronbach's Alpha)

Test Model	Listening	Reading	Speaking	Writing	Overall
Current Test	0.904	0.912	0.943	0.888	0.954
New Test	0.880	0.905	0.930	0.875	0.944

Both sets of reliability figures are very high. Reliability levels higher than 0.9 are normally only found in factually based tests consisting of multiple-choice questions (MCQ). The PTE Academic test consists of 20 different item types with a number of open-ended unconstrained response types in order to assess language skills in authentic contexts.

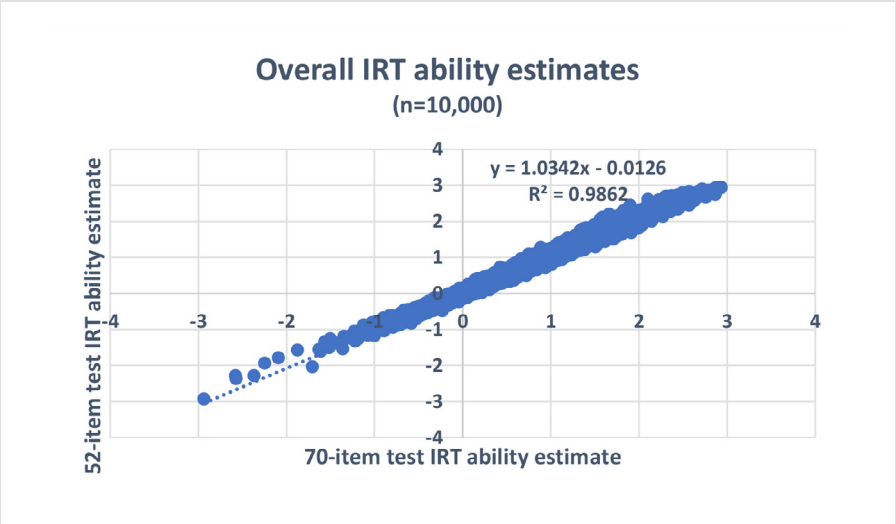
Comparison of Average Standard Error of Measurement (SEM)

Test Model	Listening	Reading	Speaking	Writing	Overall
Current Test	4.08	4.48	5.15	4.39	2.87
New Test	4.59	4.73	5.84	4.73	3.17

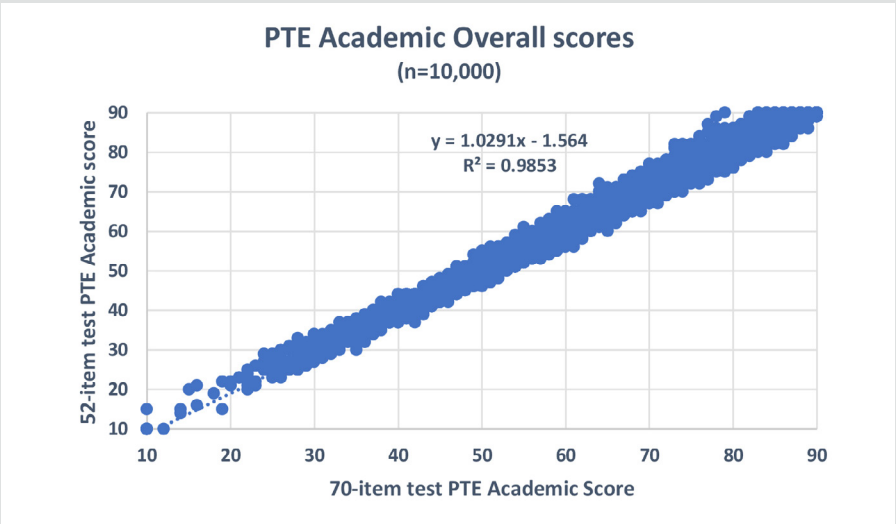
The average Standard Error of Measurement (SEM) for the current and new test model were calculated for a sample of 10,000 PTE Academic test scores. The SEM for the new test is very similar to the current test, with the overall SEM rounded to 3 GSE points. There is a maximum difference of less than one GSE point between the two models. This means that test takers and accepting institutions can be confident that PTE Academic remains the language test with the lowest error across all test providers.

Demand of the test and comparability of scores

For each potential test model, IRT ability estimates were calculated and compared to the estimates produced under the current test model. The selected model ensures that test takers would receive equivalent scores under the new model and the demand of the new test would be equivalent to the current version. Results indicate a high level of agreement between the two estimates of person abilities with a high level of explained variance ($r^2 = 0.9862$). Therefore, even though the test is shorter, the test has the same demand and test taker scores would be comparable on both test models.



PTE Academic scores are reported on a scale of 10 to 90 and are a linear transformation of the underlying IRT ability scale. The results below show the same strong relationship between the current and revised model scores on the reported score scale. The correlation between the scores across both test forms is extremely high- r> 0.99, with the explained variance $r^2 > 0.98$.



Conclusion

As a result of the analyses of performance over time, we have established that the PTE Academic can be shortened in terms of testing time, yet still retain the same level of demand. All 20 item types are still used, and each language skill is fully assessed in order to give a valid and reliable skill and overall test score. The reliability of the new test remains very high, and the Standard Error of Measurement (SEM) remains the lowest in global language testing.