Automated Scoring of English Language Skills to Achieve High Reliability, High Efficiency, and Standardization

20 August, 2018
MELTA 2018

Masa Suzuki
Director, Test Development
Pearson
先生はＡＩロボット 英語の授業で試験的実施へ 文科省

2018年8月19日 4時15分  IT・ネット
Challenges with traditional English Tests

• Listening, Reading, Grammar, Vocabulary focused

• Inadequate predictors of real-world skills

• If speaking, writing offered, human scoring introduces inconsistencies

• Limited availability/flexibility of test dates

• Weeks to get scores back
PTE ACADEMIC™

- Listening, Reading, Grammar, Vocabulary focused
  - Speaking, Writing included
- Inadequate predictors of real-world skills
  - Simulates real-life language with authentic materials
- If Speaking, Writing offered, human scoring introduces inconsistencies
  - AI-based, automated, standardized scoring
- Limited availability/flexibility of test dates
  - 360 days available for testing
- Weeks to get scores back
  - Within 5 business days
PTE Academic was created in response to higher education’s feedback for a more secure, relevant, accurate, and objective test of English.

“PTE Academic is a useful tool for ensuring that the international students we admit are able to express themselves easily in spoken and written English....”

- Rebekah Westphal, Director, Undergraduate International Admissions, Yale University
PTE ACADEMIC™

• The world’s first **fully automatically scored**, high-stakes test of academic English
• Computer-based test of **international, academic English**
• All **four** skills (Listening, Speaking, Reading, Writing)
• 3 hours of testing
• Administered at Pearson’s **certified test centers for high security**
• Objectively and consistently scored by **automated scoring systems**, including Speaking and Writing
PTE ACADEMIC

• 20 different tasks
• 11 performance-based tasks integrating multiple skills
• Assesses all English proficiency levels reliably (A1 to C2 on CEFR)
Overall score

Communicative Skills
• Speaking
• Writing
• Reading
• Listening

Enabling Skills
• Grammar
• Oral Fluency
• Pronunciation
• Spelling
• Vocabulary
• Written Discourse
PTE-A Score Report

- **Global Scale of English (GSE)**
- A granular scale between 10 and 90
- Linked to Common European Framework of Reference (CEFR)
- Scores returned within **5 business days**
Trusted around the world

- Accepted by universities in Asia
- Accepted by IIE, Fulbright.
- 100% of Australian universities, most professional associations and for all student or migration visas
- 88% of Canadian universities
- 100% of Irish universities
- 96% of UK universities
- Accepted by 2000 points of recognition in the USA
- Accepted by European education bodies and many institutions teaching English
- Accepted by major universities in the UAE
- Accepted by International Scholarships Programs, e.g. IIE, Fulbright.

For a complete listing: pearsonpte.com/accepts
Convenient

- Testing over 360 days/year.
- In over 50 countries.
- Book up to 24 hours before.
- Fast - 85% of results within 2 days.
Test Structure & Item Types
# PTE-A Test Structure

<table>
<thead>
<tr>
<th>Part</th>
<th>Content</th>
<th>Time allowed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intro</td>
<td>Introduction</td>
<td>Not timed</td>
</tr>
<tr>
<td>Part 1</td>
<td>Speaking and Writing</td>
<td>77-93 minutes</td>
</tr>
<tr>
<td>Part 2</td>
<td>Reading</td>
<td>32-41 minutes</td>
</tr>
<tr>
<td>Optional scheduled break</td>
<td></td>
<td>10 minutes</td>
</tr>
<tr>
<td>Part 3</td>
<td>Listening</td>
<td>45-57 minutes</td>
</tr>
</tbody>
</table>
# PTE-A Speaking & Writing Item Types

(7 item types)

<table>
<thead>
<tr>
<th>Item</th>
<th>Task</th>
<th>Skills assessed</th>
<th>Prompt length</th>
<th>Time to answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read aloud</td>
<td>A text appears on screen. Read the text aloud</td>
<td>reading and speaking</td>
<td>text up to 60 words</td>
<td>varies by item, depending on the length of text</td>
</tr>
<tr>
<td>Repeat sentence</td>
<td>After listening to a recording of a sentence, repeat the sentence</td>
<td>listening and speaking</td>
<td>3-9 seconds</td>
<td>15 seconds</td>
</tr>
<tr>
<td>Describe image</td>
<td>An image appears on screen. Describe the image in detail</td>
<td>speaking</td>
<td>N/A</td>
<td>40 seconds</td>
</tr>
<tr>
<td>Re-tell lecture</td>
<td>After listening to or watching a lecture, re-tell the lecture in your own words</td>
<td>listening and speaking</td>
<td>up to 90 seconds</td>
<td>40 seconds</td>
</tr>
<tr>
<td>Answer short question</td>
<td>After listening to a question, answer with a single word or a few words</td>
<td>listening and speaking</td>
<td>3-9 seconds</td>
<td>10 seconds</td>
</tr>
<tr>
<td>Summarize written text</td>
<td>After reading a text, write a one-sentence summary of the passage</td>
<td>reading and writing</td>
<td>text up to 300 words</td>
<td>10 minutes</td>
</tr>
<tr>
<td>Write essay</td>
<td>Write a 200-300 word essay on a given topic</td>
<td>writing</td>
<td>2-3 sentences</td>
<td>20 minutes</td>
</tr>
</tbody>
</table>
Describe Image

Look at the map below. In 25 seconds, please speak into the microphone and describe in detail what the map is showing. You will have 40 seconds to give your response.
Tobacco, mainly in the form of cigarettes, is one of the most widely-used drugs in the world. Over a billion adults legally smoke tobacco every day. The long term health costs are high - for smokers themselves, and for the wider community in terms of health care costs and lost productivity.

Do governments have a legitimate role to legislate to protect citizens from the harmful effects of their own decisions to smoke, or are such decisions up to the individual?
# PTE-A Reading Item Types

(5 item types)

<table>
<thead>
<tr>
<th>Item</th>
<th>Task</th>
<th>Skills assessed</th>
<th>Prompt length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple-choice, choose single answer</td>
<td>After reading a text, answer a multiple-choice question on the content or tone of the text by selecting one response</td>
<td>reading</td>
<td>text up to 110 words</td>
</tr>
<tr>
<td>Multiple-choice, choose multiple answers</td>
<td>After reading a text, answer a multiple-choice question on the content or tone of the text by selecting more than one response</td>
<td>reading</td>
<td>text up to 300 words</td>
</tr>
<tr>
<td>Re-order paragraphs</td>
<td>Several text boxes appear on screen in a random order. Put the text boxes in the correct order</td>
<td>reading</td>
<td>text up to 150 words</td>
</tr>
<tr>
<td>Reading: Fill in the blanks</td>
<td>A text appears on screen with several gaps. Drag words from the box below to fill the gaps</td>
<td>reading</td>
<td>text up to 80 words</td>
</tr>
<tr>
<td>Reading and Writing: Fill in the blanks</td>
<td>A text appears on screen with several gaps. Fill in each gap from a drop-down list of response options</td>
<td>reading and writing</td>
<td>text up to 300 words</td>
</tr>
</tbody>
</table>
PTE-A Listening Item Types
(8 item types)

<table>
<thead>
<tr>
<th>Item</th>
<th>Task</th>
<th>Skills assessed</th>
<th>Prompt length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summarize spoken text</td>
<td>After listening to a recording, write a 50-70 word summary</td>
<td>listening and writing</td>
<td>60-90 seconds</td>
</tr>
<tr>
<td>Multiple choice, choose multiple answers</td>
<td>After listening to a recording, answer a multiple-choice question on the content or tone of the recording by selecting more than one response</td>
<td>listening</td>
<td>40-90 seconds</td>
</tr>
<tr>
<td>Fill in the blanks</td>
<td>A transcript of a recording appears on screen with several gaps. After listening to the recording, type the missing word in each gap</td>
<td>listening and writing</td>
<td>30-60 seconds</td>
</tr>
<tr>
<td>Highlight correct summary</td>
<td>After listening to a recording, select the paragraph that best summarizes the recording</td>
<td>listening and reading</td>
<td>30-90 seconds</td>
</tr>
<tr>
<td>Multiple choice, choose single answer</td>
<td>After listening to a recording, answer a multiple-choice question on the content or tone of the recording by selecting one response</td>
<td>listening</td>
<td>30-60 seconds</td>
</tr>
<tr>
<td>Select missing word</td>
<td>After listening to a recording, select the missing word that completes the recording from a list of options</td>
<td>listening</td>
<td>20-70 seconds</td>
</tr>
<tr>
<td>Highlight incorrect words</td>
<td>The transcript of a recording appears on screen. While listening to the recording, identify the words in the transcript that differ from what is said</td>
<td>listening and reading</td>
<td>15-50 seconds</td>
</tr>
<tr>
<td>Write from dictation</td>
<td>After listening to a recording of a sentence, type the sentence</td>
<td>listening and writing</td>
<td>3-5 seconds</td>
</tr>
</tbody>
</table>
## Number of items on Tests

<table>
<thead>
<tr>
<th>Test</th>
<th># of Speaking items</th>
<th># of Writing items</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOEFL iBT</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>IELTS Academic</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>PTE-A</td>
<td>35</td>
<td>15</td>
</tr>
</tbody>
</table>
How do we do this?
How do we do this?
Field Testing

Initial test development involved two field tests

- Over **10,000** test takers
- **500,000** responses
- **158** countries of birth
- **126** different L1 languages

- Responses were human rated
- Training data for development of automated scoring systems
- Robust item calibration
Development of Automated Scoring Systems

System is “trained” to predict human scores

Human Transcribers and Scorers
## Traits scored

<table>
<thead>
<tr>
<th>Task</th>
<th>Traits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read Aloud</td>
<td>Content, Pronunciation, Fluency</td>
</tr>
<tr>
<td>Repeat Sentence</td>
<td>Content, Pronunciation, Fluency</td>
</tr>
<tr>
<td>Describe Image</td>
<td>Content, Pronunciation, Fluency</td>
</tr>
<tr>
<td>Retell Lecture</td>
<td>Content, Pronunciation, Fluency</td>
</tr>
<tr>
<td>Answer Short Questions</td>
<td>Content</td>
</tr>
<tr>
<td>Summarize Written Text</td>
<td>Content, Form, Grammar, Vocabulary</td>
</tr>
<tr>
<td>Write Essay</td>
<td>Content, Form, Development, Structure &amp; Coherence, Grammar, General Linguistic Range, Vocabulary Range, Spelling</td>
</tr>
<tr>
<td>Summarize Spoken Text</td>
<td>Content, Form, Grammar, Vocabulary</td>
</tr>
</tbody>
</table>
What is Transcription?

• Orthographic representations (word-for-word) of student responses

• Annotated in detail for various speech events using specific rules

  e.g.

  \[ N \text{ my *neighbor is trave(ling)- @ traveling this week} \]
Why do we need transcriptions?

To train ASR and scoring models!

Three components in ASR

1. Acoustic model (sound system)
2. Language model (grammar rules)
3. Pronunciation dictionary (Vocab and pronunciation)

Decoding
(turn speech signal into words)

$W_1, W_2, W_3 \ldots \ldots W_x$
Automated Scoring – Speaking
Multiple Aspects per Response

Parameters for content scoring

Content Scoring

Parameters for fluency

Fluency Model

Parameters for pronunciation

Pronunciation Model
Automatic Speech Recognition

Waveform

Spectrum

Words Segmentation

75-90 Words/Min

5.8 Phones/Sec
Content Scoring

Correct Answer:
“it’s supposed to rain tomorrow isn’t it”
Simplified Response Network

Example: Say what’s in the picture.
Manner Scoring

Initial silence

Segment durations compared to natives

Pause durations compared to natives

Fluency scoring model

Prediction of human FLUENCY score
Automated Scoring – Writing
Latent Semantic Analysis (LSA)

<table>
<thead>
<tr>
<th>Key Word</th>
<th>LSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doctor—Doctor</td>
<td>1.0</td>
</tr>
<tr>
<td>Doctor—Physician</td>
<td>0.8</td>
</tr>
<tr>
<td>Doctor—Surgeon</td>
<td>0.7</td>
</tr>
</tbody>
</table>

“Surgery is often performed by a team of doctors.”

“On many occasions, several physicians are involved in an operation.”

Key Word = 0

LSA = 0.73
Semantic Space for Essay Scoring

300+ semantic dimensions

New Essay Score?  
Pre-Scored ‘6’  
Pre-Scored ‘2’  
Pre-Scored ‘4’
Essay scoring process

Essay Score

- Content
- Style
- Mechanics
- Scoring confidence
- Similarity to expert scored essays
- Coherence
- Grammar
- Off-topic detection
Validation
Development

System is “trained” to predict human scores

Validation

Expert human ratings

Very highly correlated

Machine scores
# Test Reliability

<table>
<thead>
<tr>
<th></th>
<th>PTE Academic</th>
<th>IELTS</th>
<th>TOEFL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.97</td>
<td>0.96</td>
<td>0.94</td>
</tr>
<tr>
<td>Reading</td>
<td>0.92</td>
<td>0.90</td>
<td>0.85</td>
</tr>
<tr>
<td>Listening</td>
<td>0.91</td>
<td>0.91</td>
<td>0.85</td>
</tr>
<tr>
<td>Writing</td>
<td>0.91</td>
<td>0.81-0.90</td>
<td>0.74</td>
</tr>
<tr>
<td>Speaking</td>
<td>0.91</td>
<td>0.83-0.86</td>
<td>0.88</td>
</tr>
</tbody>
</table>

- Acceptable
- Good
- Very Good
PTE-A Speaking Scores – Accuracy

Expert human ratings 

Very highly correlated

Machine scores

<table>
<thead>
<tr>
<th>Machine-Human Correlation (N=158)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pronunciation</td>
<td>0.81</td>
</tr>
<tr>
<td>Fluency</td>
<td>0.82</td>
</tr>
<tr>
<td>Content</td>
<td>0.92</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>0.90</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.95</td>
</tr>
<tr>
<td>Overall</td>
<td>0.96</td>
</tr>
</tbody>
</table>
Benefits of automated scoring

Automated scoring systems

- Standardized scoring
- Consistent scoring
- Speed of scoring
- Objective, bias-free measurement
- Data-driven models from 10,000+ candidates
- Accumulation of measures from multiple expert raters
Automated Language Tests

Corporate
• Recruitment screening
• Training & leadership programs
• Aviation & transportation

Education
• Teacher/TA certification
• English Learners - identification, certification, progress monitoring
• University admissions/qualification

Government
• Immigration
• On-the-job certification
The story of Mr. W with Versant
Super Testing Man
Mr. W
Have you ever wondered whether you might be able to improve your test score if you could take as many Versant Tests as you like?

In a four-day period, Mr. W. took the Versant English speaking test 46 times!
Mr. W.’s Overall Score results over all valid test occasions
Investigation of Mr. W’s tests revealed two interesting facts.

1. First four tests were taken between 11pm and 1:30am on a Thursday night. Mr. W’s scores did not improve in the early hours of the morning. The remainder of Mr. W’s test occasions, all 42 of them, were taken between the more sensible hours of 10.30am and 10pm.

2. The final two tests (the ones with the highest scores) were not Mr. W’s at all! Listening to the tests in the Versant database, it was apparent that their voices are different.

Thus, 6 tests are suspect!
The Full Picture of Reliability

If we remove these 6 tests from the investigation, then Mr. W’s score pattern is basically flat and stable.

On the Common European Framework of Reference (CEFR), the Versant score range for B1 is 47-57. Mr. W is evaluated by Versant as a B1 speaker of English every single time.

Mr. W’s average score between 50 - 57
And to ‘Mr. W’ we say:

We salute your perseverance! At 5.30pm on a **Friday evening you took Versant and scored 55**. Some fifty hours later, after thirty-nine Versant tests and a twelve-hour marathon of test taking, **your score increased to ... 56!**

---

**FRIDAY**

**SCORE 55**

**MONDAY**

**SCORE 56**

55 Hours Later 39 Versant Tests

Twelve-hour marathon of test taking
“Despite its name, there is nothing “artificial” about this technology — it is made by humans, intended to behave like humans and affects humans.

.... No technology is more reflective of its creators than A.I. It has been said that there are no “machine” values at all, in fact; machine values are human values. A human-centered approach to A.I. means these machines don’t have to be our competitors, but partners in securing our well-being.”

Prof. Fei-Fei Li, at Stanford University
(March 7, 2018, NY Times)
ALWAYS LEARNING
## Content Rubric

<table>
<thead>
<tr>
<th>Communicative skills</th>
<th>Speaking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enabling skills and other traits scored</td>
<td>Content:</td>
</tr>
<tr>
<td></td>
<td>5 Describes all elements of the image and their relationships, possible development and conclusion or implications</td>
</tr>
<tr>
<td></td>
<td>4 Describes all the key elements of the image and their relations, referring to their implications or conclusions</td>
</tr>
<tr>
<td></td>
<td>3 Deals with most key elements of the image and refers to their implications or conclusions</td>
</tr>
<tr>
<td></td>
<td>2 Deals with only one key element in the image and refers to an implication or conclusion. Shows basic understanding of several core elements of the image</td>
</tr>
<tr>
<td></td>
<td>1 Describes some basic elements of the image, but does not make clear their interrelations or implications</td>
</tr>
<tr>
<td></td>
<td>0 Mentions some disjointed elements of the presentation</td>
</tr>
</tbody>
</table>
# Pronunciation

<table>
<thead>
<tr>
<th>Pronunciation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>5 Native-like</strong></td>
<td>All vowels and consonants are produced in a manner that is easily understood by regular speakers of the language. The speaker uses assimilation and deletions appropriate to continuous speech. Stress is placed correctly in all words and sentence-level stress is fully appropriate.</td>
</tr>
<tr>
<td><strong>4 Advanced</strong></td>
<td>Vowels and consonants are pronounced clearly and unambiguously. A few minor consonant, vowel or stress distortions do not affect intelligibility. All words are easily understandable. A few consonants or consonant sequences may be distorted. Stress is placed correctly on all common words, and sentence level stress is reasonable.</td>
</tr>
<tr>
<td><strong>3 Good</strong></td>
<td>Most vowels and consonants are pronounced correctly. Some consistent errors might make a few words unclear. A few consonants in certain contexts may be regularly distorted, omitted or mispronounced. Stress-dependent vowel reduction may occur on a few words.</td>
</tr>
<tr>
<td><strong>2 Intermediate</strong></td>
<td>Some consonants and vowels are consistently mispronounced in a non-native like manner. At least 2/3 of speech is intelligible, but listeners might need to adjust to the accent. Some consonants are regularly omitted, and consonant sequences may be simplified. Stress may be placed incorrectly on some words or be unclear.</td>
</tr>
<tr>
<td><strong>1 Intrusive</strong></td>
<td>Many consonants and vowels are mispronounced, resulting in a strong intrusive foreign accent. Listeners may have difficulty understanding about 1/3 of the words. Many consonants may be distorted or omitted. Consonant sequences may be non-English. Stress is placed in a non-English manner; unstressed words may be reduced or omitted and a few syllables added or missed.</td>
</tr>
<tr>
<td><strong>0 Non-English</strong></td>
<td>Pronunciation seems completely characteristic of another language. Many consonants and vowels are mispronounced, misordered or omitted. Listeners may find more than 1/2 of the speech unintelligible. Stressed and unstressed syllables are realized in a non-English manner. Several words may have the wrong number of syllables.</td>
</tr>
</tbody>
</table>
# Content

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>The response provides a good summary of the text.</td>
</tr>
<tr>
<td>1</td>
<td>The response provides a fair summary of the text, but misses one or two aspects.</td>
</tr>
<tr>
<td>0</td>
<td>The response omits or misrepresents the main issue(s) dealt with in the text.</td>
</tr>
<tr>
<td>9</td>
<td>There is no response, response is not English or irrelevant</td>
</tr>
</tbody>
</table>
### Development, Structure & Coherence

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>The essay shows a good development and logical structure</td>
</tr>
<tr>
<td>1</td>
<td>The essay is less well structured, some elements or paragraphs seem poorly linked</td>
</tr>
<tr>
<td>0</td>
<td>The essay lacks coherence, mainly consists of lists or loose elements</td>
</tr>
<tr>
<td>9</td>
<td>There is no response, response is not English or irrelevant</td>
</tr>
</tbody>
</table>
Auto-scoring can assess these skills

**Written Scoring**
- Word choice
- Grammar & Mechanics
- Progression of ideas
- Organization
- Style, Tone
- Paragraph structure
- Development, Coherence
- Point of view
- Task completion

**Spoken Scoring**
- Sentence Mastery
- Content
- Vocabulary
- Accuracy
- Pronunciation
- Intonation
- Fluency
- Expressiveness
- Pragmatics