Re-defining the High-stakes Test of English with Automated Scoring Technology

21 July, 2017 @ VUS TESOL
Masanori Suzuki
Director, Test Development
Pearson
What’s common here?
Overview

1. General introduction of PTE Academic
2. Brief Overview of Test Development Process
3. Automated Scoring for Writing
4. Automated Scoring for Speaking
please translate these roman numerals mcmxcviii thank you
“I asked my nan why she used 'please' and 'thank you' and it seemed she thinks that there is someone - a physical person - at Google's headquarters who looks after the searches.”
Educational Technology

Instruction

Assessment

High-stakes
Summative

Low-stakes
Formative

Language Assessment
1970’s – Technology!
Bubble cards & readers

• Multi-choice format
• Passive skills

• Efficiency in scoring
• Objectivity in scoring
Pearson Test of English (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 (c.f. TOEFL = 4 hours)
- Administered at Pearson’s **certified test centers for high security**
- Objectively and consistently scored **by automated scoring systems**, including Speaking and Writing
Pearson Test of English (PTE) - Academic

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

Unrivaled security measures across all test centers

- Palm-vein authentication
- Digital photographs
- Electronic signatures
- Video and audio monitoring
- Paperless testing and results
- Randomized test forms
- Secure administration
- Data forensics
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
Overall score

Communicative Skills
- Speaking
- Writing
- Reading
- Listening

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

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

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>
## 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>
Relevant

How well a test reflects the real life demands of study is an important quality of a test of academic English.

**Authenticity** is ensured in PTE-A by the use of **genuine academic test content**, setting **academically relevant tasks**, and by **measuring skills in an integrated way**.
Relevant & Objective

Assessments should provide test takers with the confidence to succeed.

- Genuine academic content so your students are better prepared to use English at your institution.
- Integrated tasks that test more than one language skill reflecting the combinations of skills students need.
- Including international and non-native English making the test more appropriate test for modern global universities.

Relevant & Objective Assessments should provide test takers with the confidence to succeed.
How did we develop the test?
Test Development Process

- Test Spec
- Item Writing
- External Review I
- External Review II
- Field Testing
  - Test Form Check
  - Editorial Review
  - Internal Review
- Native
- L2 speakers
- Transcription
- Human Rating
  - Automated Scoring Dev
  - Psychometric Analysis
  - Live Test

Native and L2 speakers contribute to the transcribed material, which is then reviewed and rated by a human. The live test provides additional validation.
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

- Training data for development of automated scoring systems
- Robust item calibration
Automated Scoring –
Where is it used now?
Automated **Spoken Tests**

**Corporate**
- Recruitment screening
- Training & leadership programs
- Aviation & transportation

**Education**
- Teacher/TA certification
- English Language Learners - placement

**Government**
- Immigration
- On-the-job certification
## Automated Spoken Tests

<table>
<thead>
<tr>
<th>Automated Test</th>
<th>Correlation to human raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>.97</td>
</tr>
<tr>
<td>Dutch</td>
<td>.93</td>
</tr>
<tr>
<td>Arabic</td>
<td>.98</td>
</tr>
<tr>
<td>French</td>
<td>.97</td>
</tr>
<tr>
<td>Chinese</td>
<td>.96</td>
</tr>
<tr>
<td>English</td>
<td>.97</td>
</tr>
</tbody>
</table>

### Corporate
- Recruitment screening
- Training & leadership programs
- Aviation & transportation

### Education
- Teacher/TA certification
- English Language Learners - placement

### Government
- Immigration
- On-the-job certification
Automated **Written Tests**

**State Assessments**
- Maryland, Virginia
- PARCC: Partnership for Assessment of Readiness for College and Careers
- ACT Aspire: College and Career Readiness

**Writing Practice:** throughout US schools & districts
- Prentice Hall: textbook companion
- HMH Riverside: textbook companion
- Pearson MyLabs: Higher Ed courseware
- Pearson WriteToLearn

**Higher Ed Placement/Evaluation**
- College Board’s ACCUPLACER (college placement)
- Council for Aid to Education (national report card for colleges)
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
Automated Scoring – Writing
Why automated scoring?

• Consistent application of rubrics by many raters
• Consistent application of rubrics of the same rater
• Consistent application of rubrics across time
• Standardized presentation of questions
• Cross-contamination among different traits
• Potential raters bias due to gender, culture, ethnicity, appearance, accents, etc.
Development of Automated Scoring Systems

System is “trained” to predict human scores.
Intelligent Essay Assessor (IEA)

Trained human raters rate essays on aspects defined in scoring rubrics: Content, Style, Mechanics

Content
• Semantic analysis measures of similarity to prescored responses, ideas, examples, ....

Style / Coherence
• Appropriate word choice, word and sentence flow, coherence

Mechanics
• Grammar, word usage, punctuation, spelling, ...
Latent Semantic Analysis (LSA)

• LSA reads lots of text
  – *For science, it reads lots of science textbooks*
• Learns what words mean and how they relate to each other
  – *Learns the concepts, not just the vocabulary*
• Result is a “Semantic Space”
  – Every word represented as a vector

Essays are compared to each other in semantic space as similarity is used to derive measures of quality as determined by human raters.
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.”

LSA = 0.73
Essay Scoring

300+ semantic dimensions

New Essay Score?

Pre-_scored ‘6’

Pre-scored ‘4’

Pre-scored ‘2’
Essay scoring process

Essay Score

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

Coherence
Other IEA features

- Detects Off-topic or highly unusual essays
- Detects if the IEA may not score an essay well
- Detects larding of big words, non-standard language constructions, swear words, too long, too short ...
- Uses non-coachable measures
  - No counts of total words, syllables, characters, etc.
  - No trigger surface features: “thus”, “therefore”
- Plagiarism
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?
<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>
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 everyday. Recently, it is not only the adult. Even the high school students or college students smoke just because they want to know how it feels. It is also not limited by gender. Lots of women are smokers. Even the old people still smoke, as if they do not care about their healthy. Become a smoker is like make someone just care about the good feeling of smoking and makes them to forget the risks they will face in the future.

The long term health costs are high - for smokers themselves, and for the wider community in terms of health care costs and lost productivity. The worst risk that the smokers will face is lung cancer, which can cause death. The governments have a legitimate role to legislate to protect citizens from the harmful effects of their own decisions to smoke. For example they make rule about no smoking area, in the street, and public place. But it also the decisions of each individual whether they want to continue their life as a smoker and take all the risk, or stop and learn to life healthier.
## Score Comparison

<table>
<thead>
<tr>
<th>Trait</th>
<th>Human 1</th>
<th>Human 2</th>
<th>Adjudicator</th>
<th>IEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>2</td>
<td>2</td>
<td></td>
<td>1.80</td>
</tr>
<tr>
<td>Development, Structure &amp; Coherence</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1.35</td>
</tr>
<tr>
<td>General Linguistic Range</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1.03</td>
</tr>
<tr>
<td>Grammar Usage &amp; Mechanics</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1.07</td>
</tr>
<tr>
<td>Vocabulary Range</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0.93</td>
</tr>
</tbody>
</table>
Automated Scoring – Speaking
Automatic Speech Recognition

Waveform

Spectrum

Words Segmentation

75-90 Words/Min

5.8 Phones/Sec
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
Example: Native Speaker

REPEAT: New York City is famous for its ethnic diversity.

“New York city is famous for its ethnic diversity”

Pronunciation: 8.7
Fluency: 8.1
Accuracy: 0 word errors
Example: English Learner

REPEAT: New York City is famous for its ethnic diversity.

“New York city is famous for its ethnic ethnic diversity”

Pronunciation: 5.9  
Fluency: 3.3  
Accuracy: 1 word error (insertion)
Performance Comparison

Pronunciation

Fluency

Accuracy

3.026 seconds
Native speaker

5.502 seconds
Learner
Multiple Aspects per Response

Parameters for content scoring

Content Scoring

Parameters for fluency

Fluency Model

Parameters for pronunciation

Pronunciation Model
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>

![Score Scale](image)
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>
Automated scoring

Automated scoring systems

- Standardized scoring
- Speed of scoring
- Objective, bias-free measurement
- Data-driven models from 10,000 candidates
- Accumulation of measures from multiple expert raters
Limitations to Automated Systems

1. They are scoring systems, rather than corrective feedback systems

2. Automated tests require a test design which plays to the strengths and limitations of automated scoring

3. Difficult to adapt items in real-time

1. Its goal is to predict human scores, not understand the response entirely; difficult to detect nuance, shades of meaning
In summary, PTE Academic provides ...

1. **Convenient**
   - Testing over 360 days/year.
   - In over 50 countries.

2. **Secure**
   - Advanced multi layer security measures.

3. **Accurate**
   - Computer based marking ensures impartial, accurate marks.

4. **Relevant & Objective**
   - Uses genuine academic content and integrated tasks testing multiple skills.

- Book up to 24 hours before.
- Most secure tests.
- Confidence in results.

- Institutions can trust the English ability of students.
- Reflects the use of language students need.

Fast - 85% of results within 2 days.
ALWAYS LEARNING