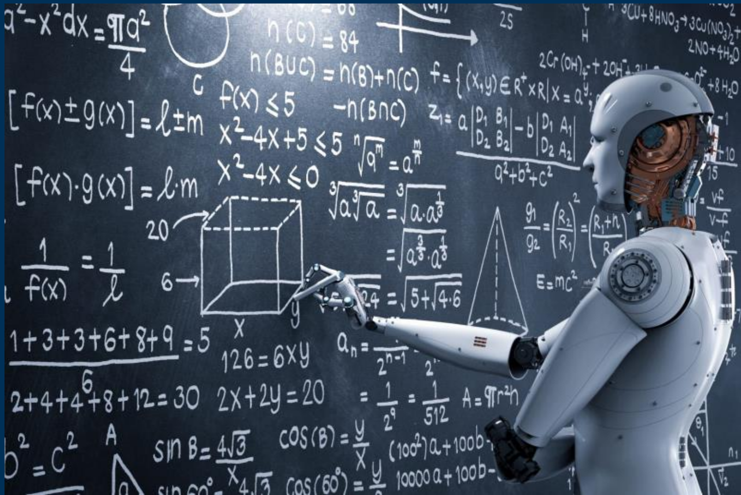
Pearson

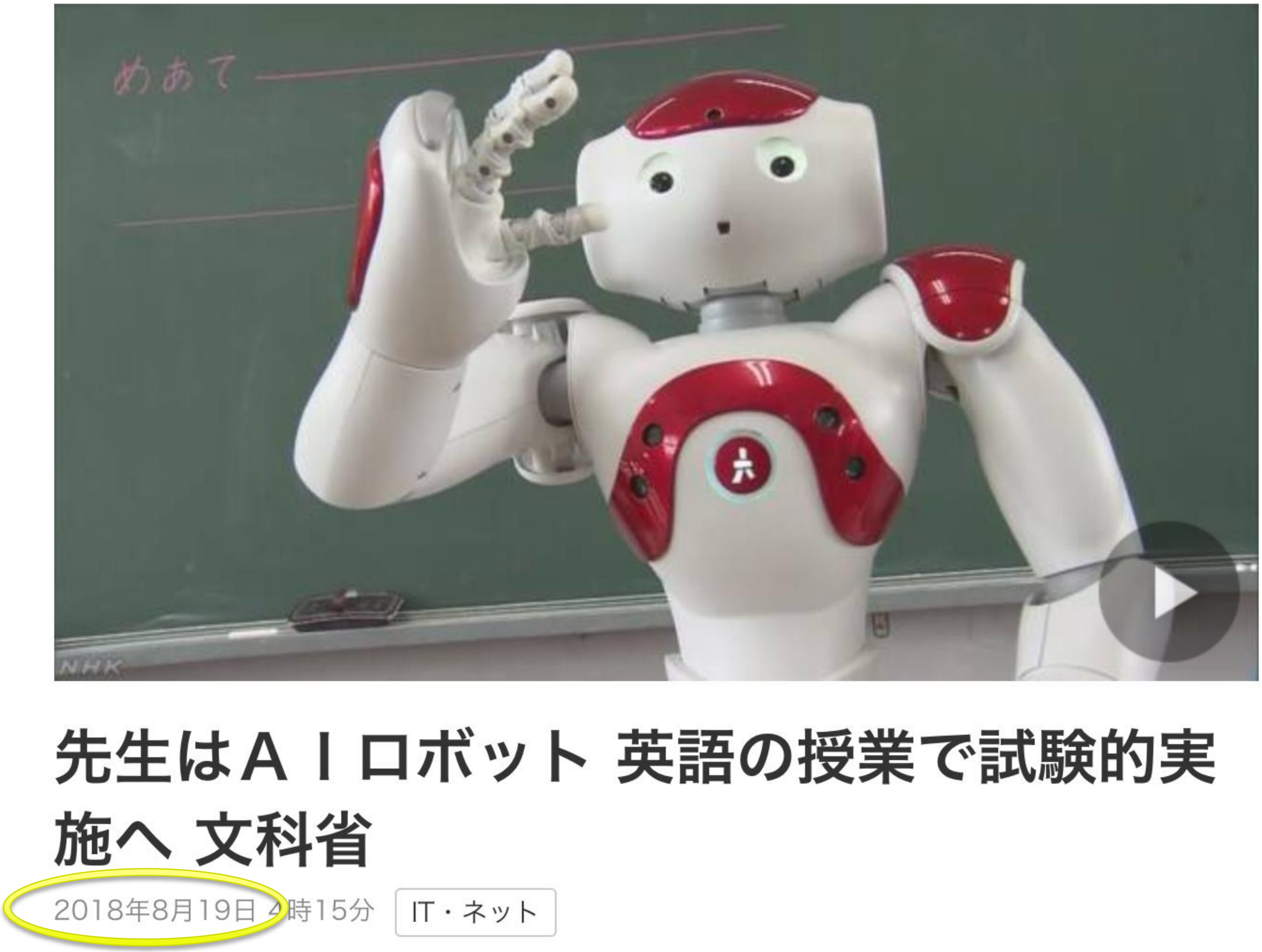
PTE ACADEMIC™

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™

Pearson

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

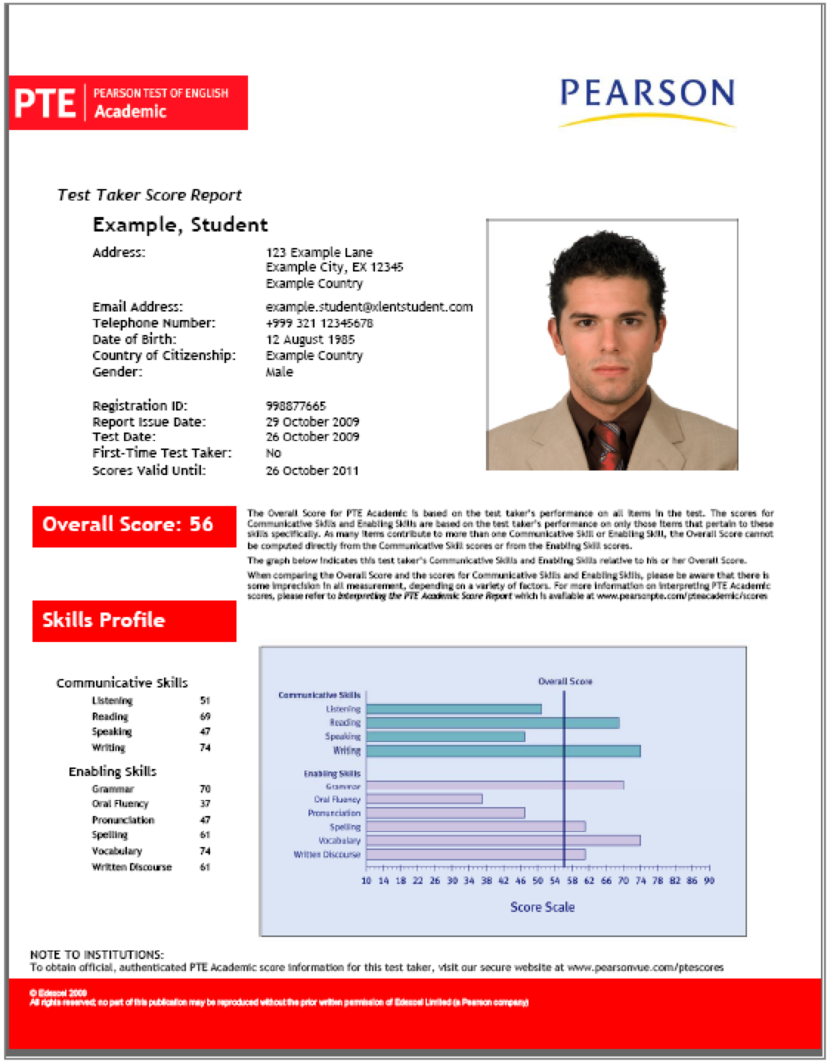
PTE ACADEMIC™

- 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)

PTE ACADEMIC™



PTE | PEARSON TEST OF ENGLISH Academic

PEARSON

Test Taker Score Report

Example, Student

Address:	123 Example Lane Example City, EX 12345 Example Country
Email Address:	example.student@xlentstudent.com
Telephone Number:	+999 321 12345678
Date of Birth:	12 August 1985
Country of Citizenship:	Example Country
Gender:	Male
Registration ID:	998877665
Report Issue Date:	29 October 2009
Test Date:	26 October 2009
First-Time Test Taker:	No
Scores Valid Until:	26 October 2011

Overall Score: 56

The Overall Score for PTE Academic is based on the test taker's performance on all items in the test. The scores for Communicative Skills and Enabling Skills are based on the test taker's performance on only those items that pertain to these skills specifically. As many items contribute to more than one Communicative Skill or Enabling Skill, the Overall Score cannot be computed directly from the Communicative Skill scores or from the Enabling Skill scores.

The graph below indicates this test taker's Communicative Skills and Enabling Skills relative to his or her Overall Score.

When comparing the Overall Score and the scores for Communicative Skills and Enabling Skills, please be aware that there is some imprecision in all measurement, depending on a variety of factors. For more information on interpreting PTE Academic scores, please refer to *Interpreting the PTE Academic Score Report* which is available at www.pearsonpte.com/pteacademic/scores

Skills Profile

Communicative Skills	
Listening	51
Reading	69
Speaking	47
Writing	74
Enabling Skills	
Grammar	70
Oral Fluency	37
Pronunciation	47
Spelling	61
Vocabulary	74
Written Discourse	61

NOTE TO INSTITUTIONS:
To obtain official, authenticated PTE Academic score information for this test taker, visit our secure website at www.pearsonvue.com/ptescores

Overall score

Communicative Skills

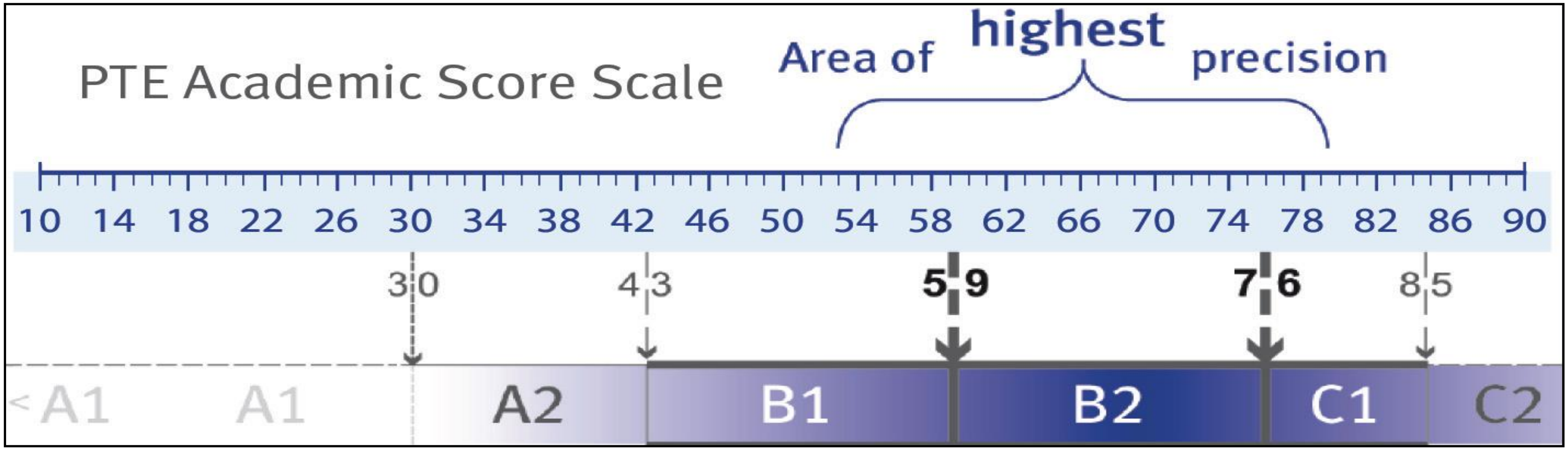
- Speaking
- Writing
- Reading
- Listening

Enabling Skills

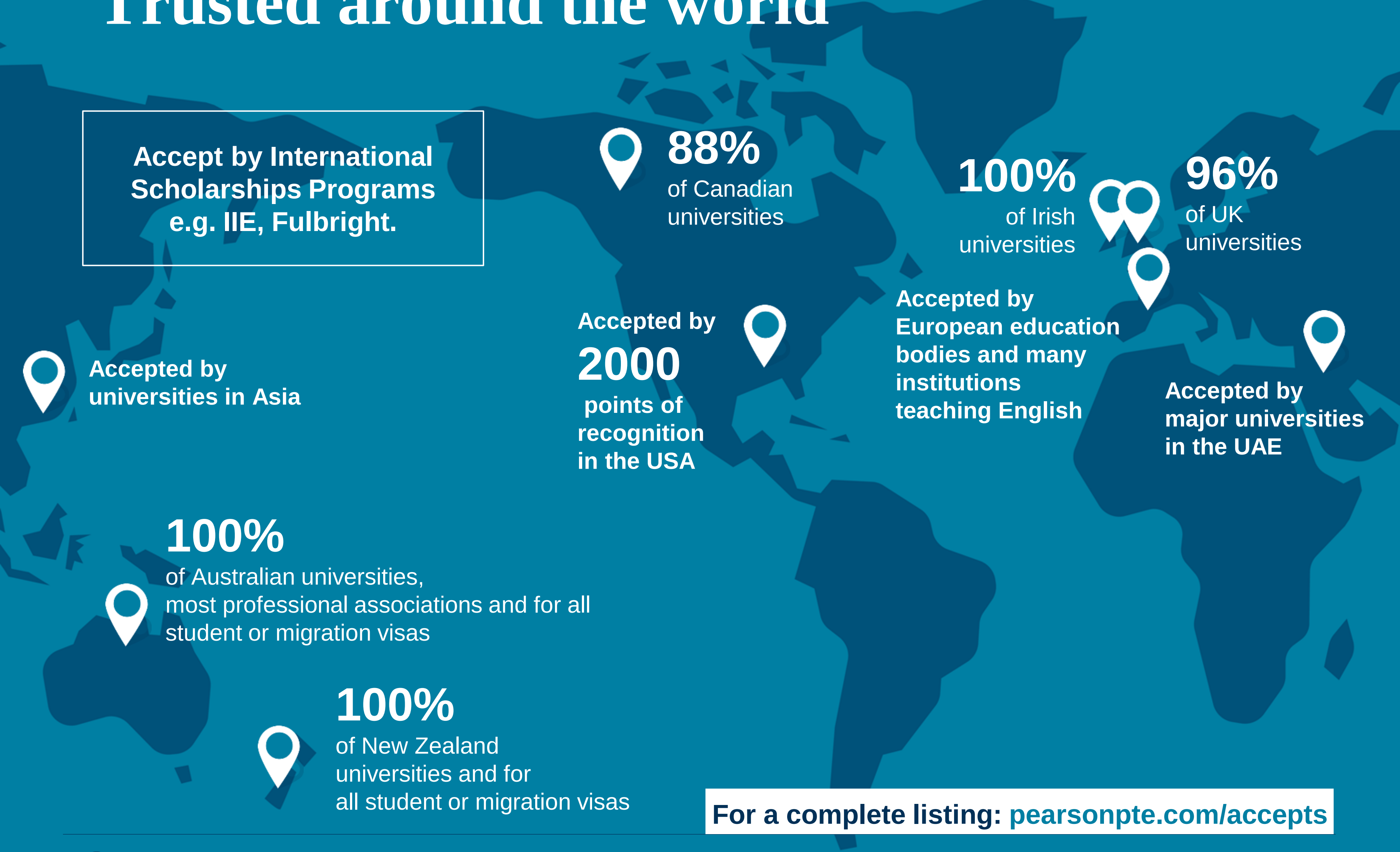
- Grammar
- Oral Fluency
- Pronunciation
- Spelling
- Vocabulary
- Written Discourse

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



Accept by International Scholarships Programs e.g. IIE, Fulbright.

Accepted by universities in Asia

88%
of Canadian universities

Accepted by
2000
points of recognition in the USA

100%
of Irish universities

96%
of UK universities

Accepted by European education bodies and many institutions teaching English

Accepted by major universities in the UAE

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

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.

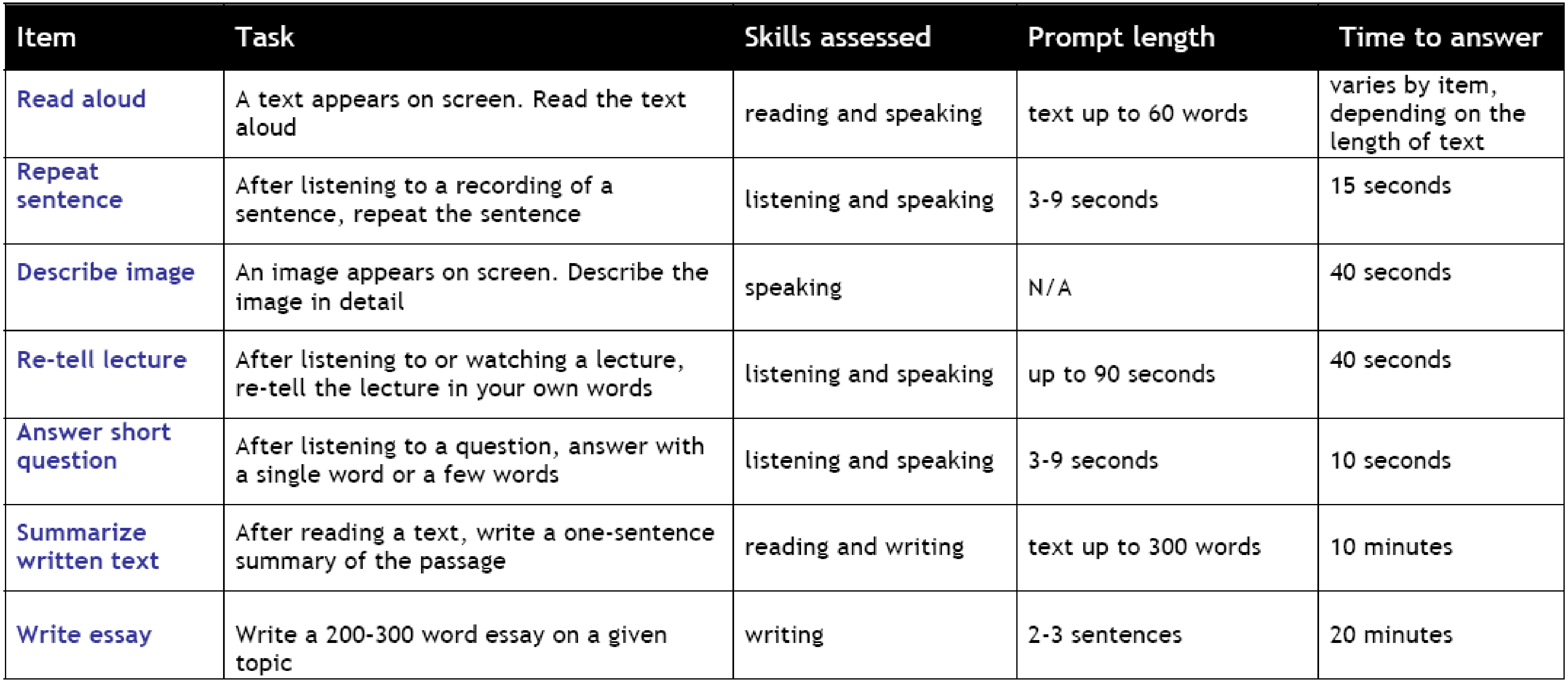
Test Structure & Item Types

Pearson

PTE-A Test Structure

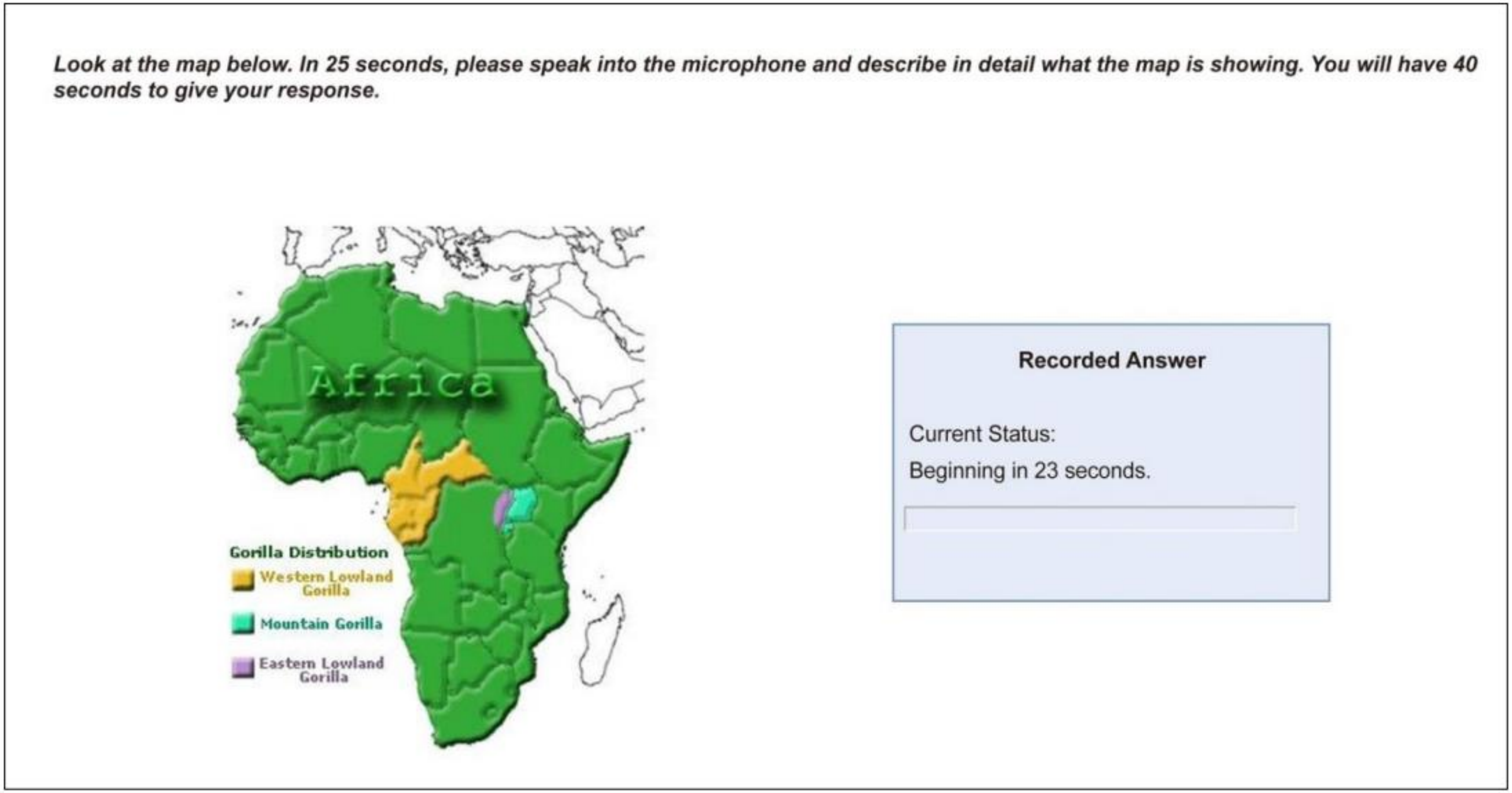
Part	Content	Time allowed
Intro	Introduction	Not timed
Part 1	Speaking and Writing	77-93 minutes
Part 2	Reading	32-41 minutes
Optional scheduled break		10 minutes
Part 3	Listening	45-57 minutes

PTE-A Speaking & Writing Item Types (7 item types)



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

Describe Image



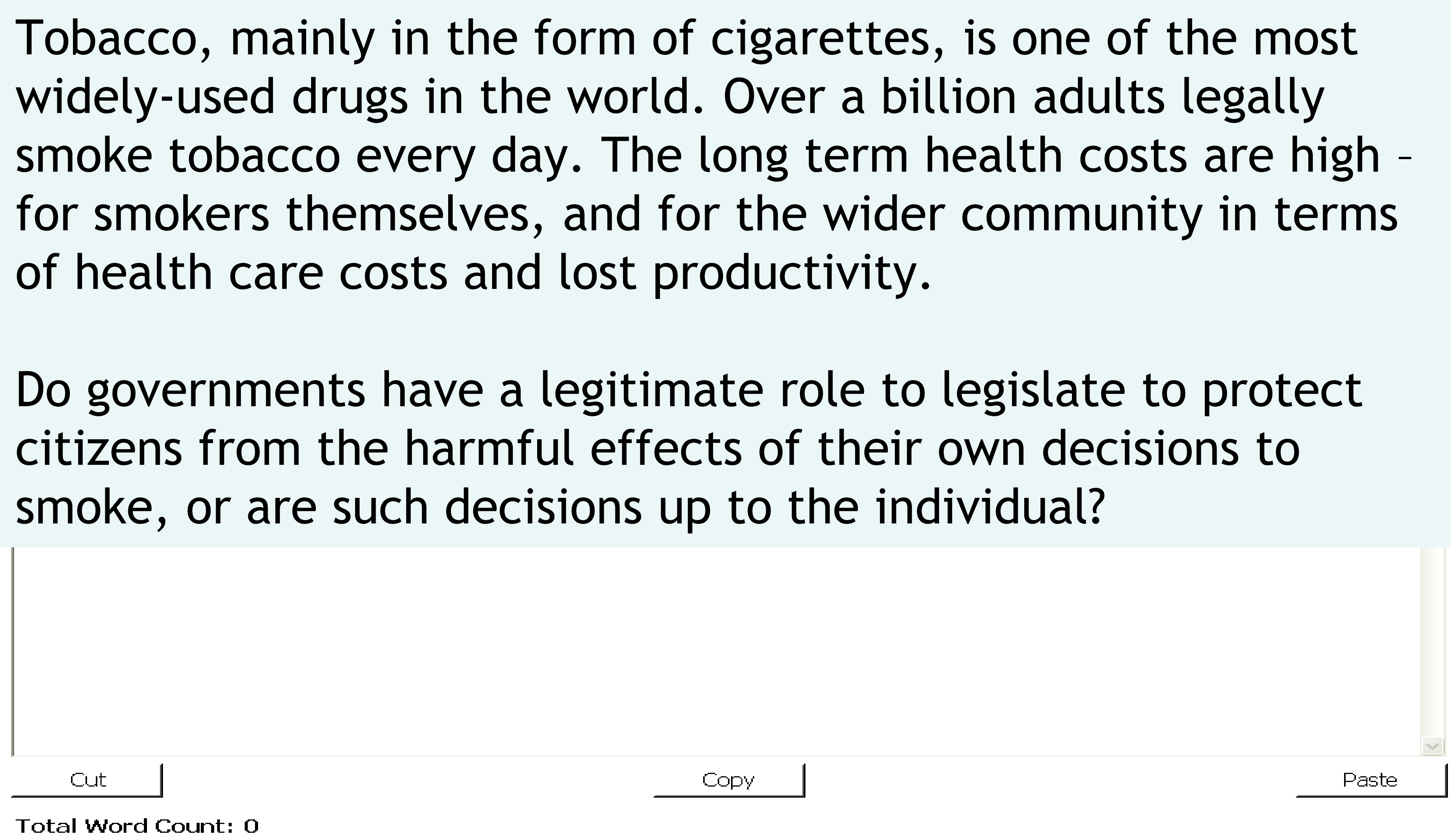
Look at the map below. In 25 seconds, please speak into the microphone and describe in detail what the map is showing. You will have 40 seconds to give your response.

Recorded Answer

Current Status:

Beginning in 23 seconds.

Write Essay

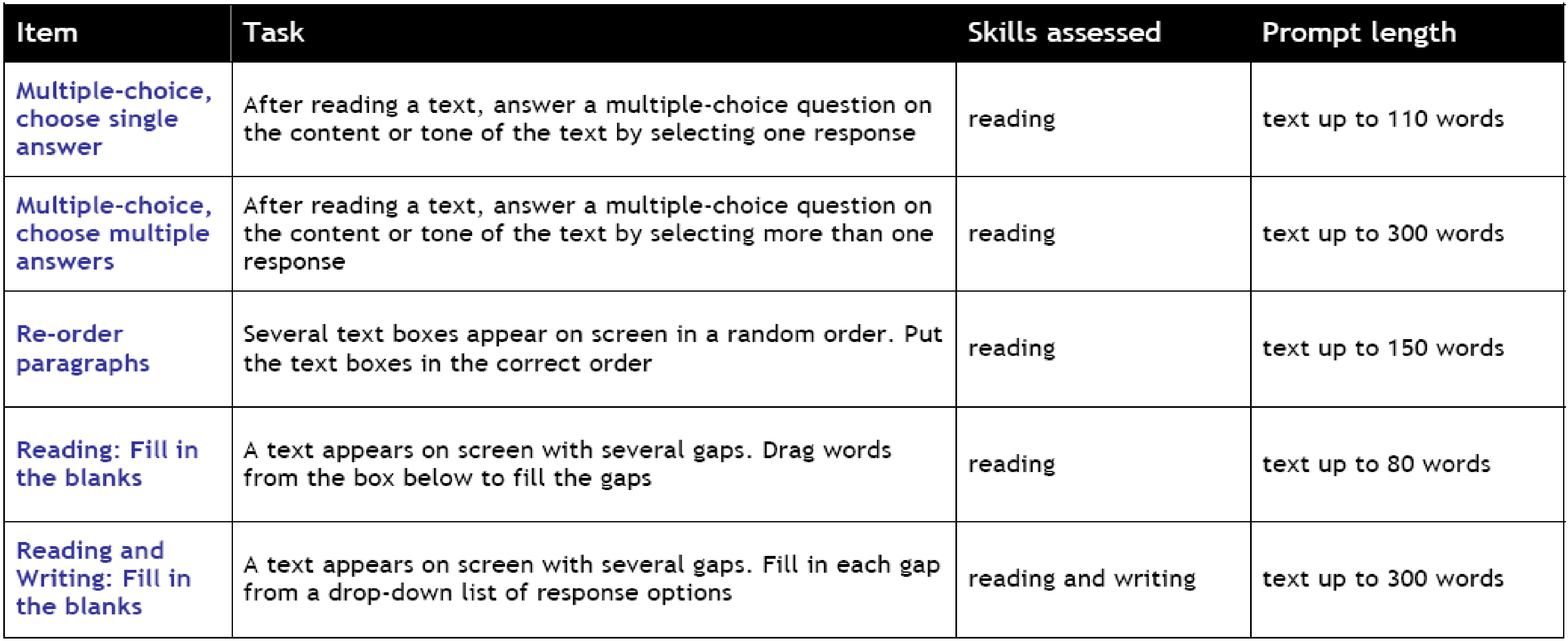


Tobacco, mainly in the form of cigarettes, is one of the most widely-used drugs in the world. Over a billion adults legally smoke tobacco every day. The long term health costs are high – for smokers themselves, and for the wider community in terms of health care costs and lost productivity.

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?

Total Word Count: 0

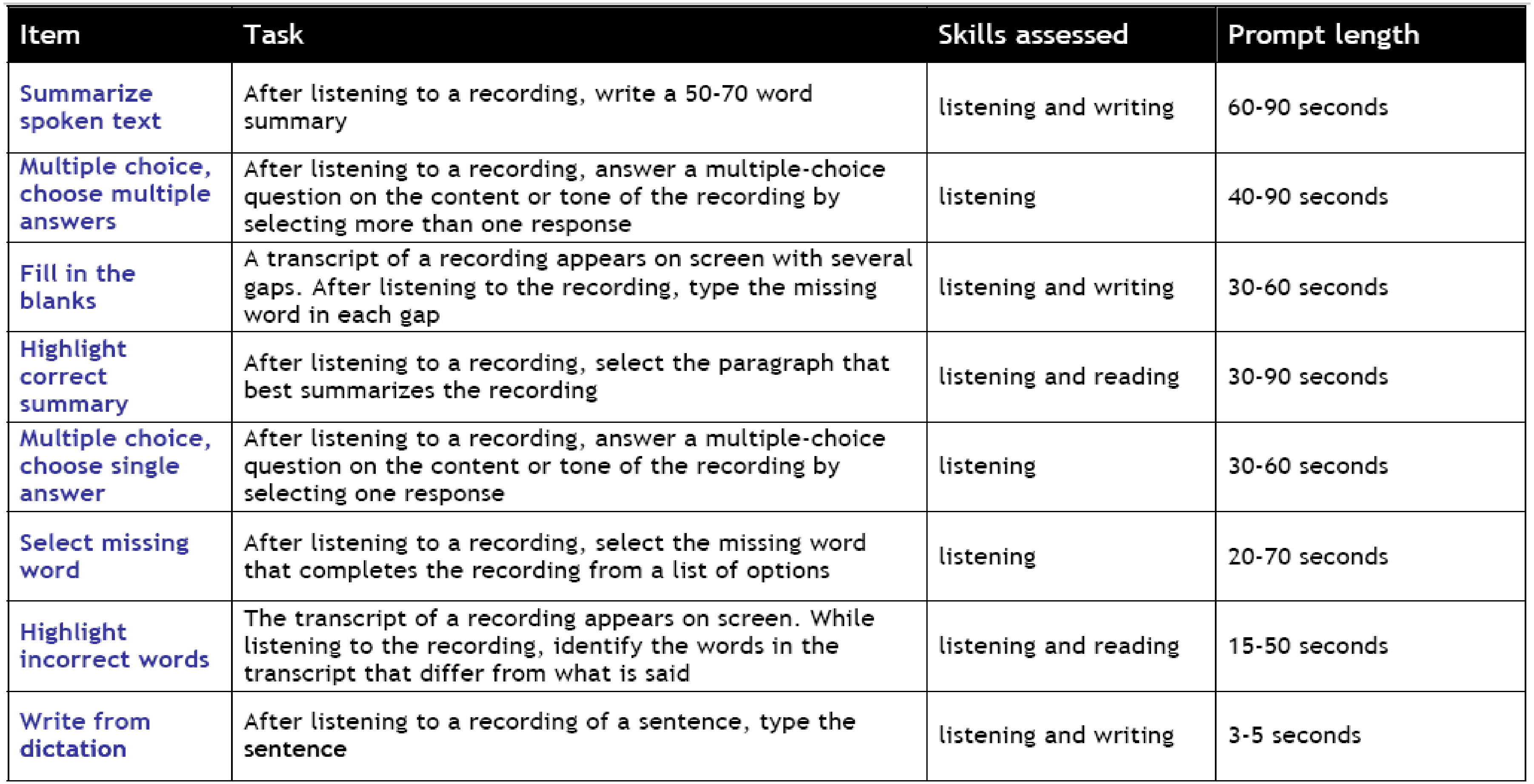
PTE-A Reading Item Types
(5 item types)



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

PTE-A Listening Item Types

(8 item types)

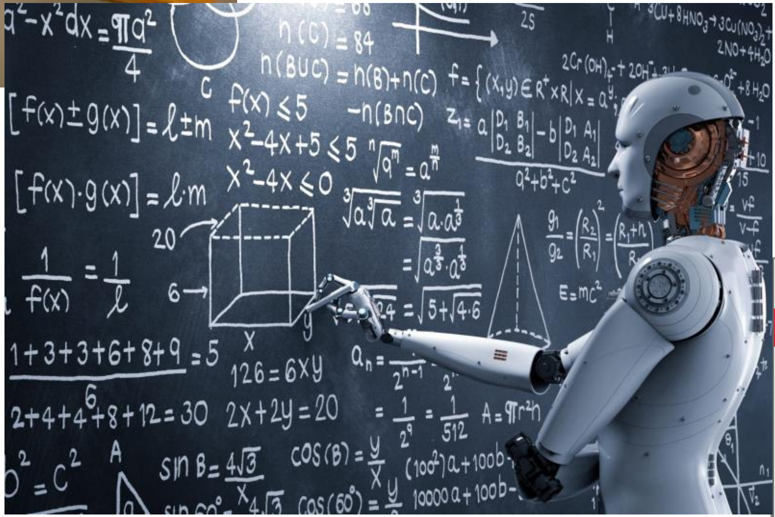


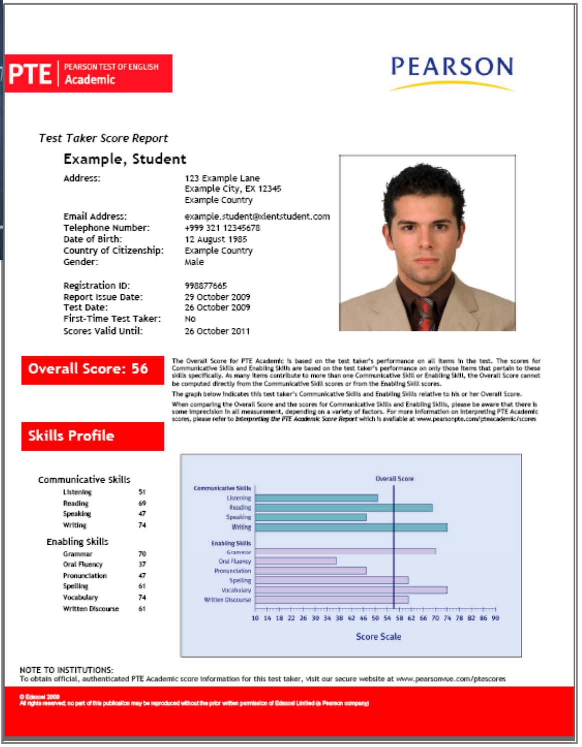
Item	Task	Skills assessed	Prompt length
Summarize spoken text	After listening to a recording, write a 50-70 word summary	listening and writing	60-90 seconds
Multiple choice, choose multiple answers	After listening to a recording, answer a multiple-choice question on the content or tone of the recording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

Number of items on Tests

Test	# of Speaking items	# of Writing items
TOEFL iBT	6	2
IELTS Academic	3	2
PTE-A	35	15

How do we do this?





PTE | PEARSON TEST OF ENGLISH Academic

PEARSON

Test Taker Score Report

Example, Student

Address:	123 Example Lane Example City, EX 12345 Example Country
Email Address:	example.student@xlentstudent.com
Telephone Number:	+999 321 12345678
Date of Birth:	12 August 1985
Country of Citizenship:	Example Country
Gender:	Male
Registration ID:	998877665
Report Issue Date:	29 October 2009
Test Date:	26 October 2009
First-Time Test Taker:	No
Scores Valid Until:	26 October 2011

Overall Score: 56

The Overall Score for PTE Academic is based on the test taker's performance on all items in the test. The scores for Communicative Skills and Enabling Skills are based on the test taker's performance on only those items that pertain to these skills specifically. As many items contribute to more than one Communicative Skill or Enabling Skill, the Overall Score cannot be computed directly from the Communicative Skill scores or from the Enabling Skill scores.

The graph below indicates this test taker's Communicative Skills and Enabling Skills relative to his or her Overall Score.

When comparing the Overall Score and the scores for Communicative Skills and Enabling Skills, please be aware that there is some imprecision in all measurement, depending on a variety of factors. For more information on interpreting PTE Academic scores, please refer to ***Interpreting the PTE Academic Score Report*** which is available at www.pearsonpte.com/pteacademic/scores

Skills Profile

Communicative Skills	
Listening	51
Reading	69
Speaking	47
Writing	74
Enabling Skills	
Grammar	70
Oral Fluency	37
Pronunciation	47
Spelling	61
Vocabulary	74
Written Discourse	61

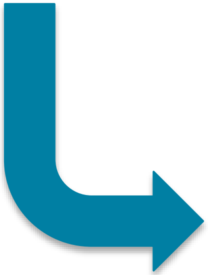
NOTE TO INSTITUTIONS:
To obtain official, authenticated PTE Academic score information for this test taker, visit our secure website at www.pearsonvue.com/ptescores

How do we do this?

Field Testing

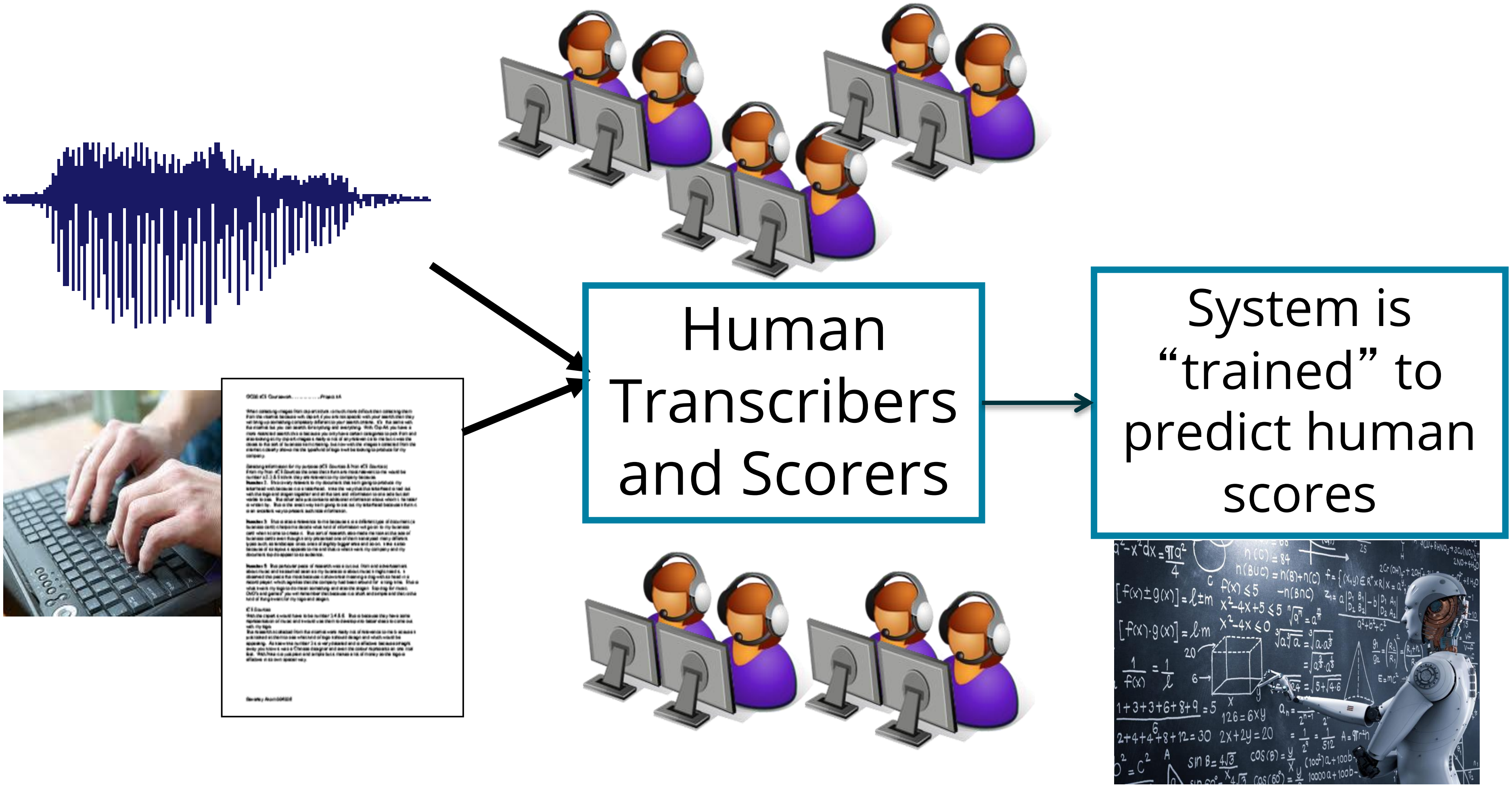
Initial test development involved two field tests

- 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



Human Transcribers and Scorers

System is “trained” to predict human scores

Traits scored

Task	Traits
Read Aloud	Content, Pronunciation, Fluency
Repeat Sentence	Content, Pronunciation, Fluency
Describe Image	Content, Pronunciation, Fluency
Retell Lecture	Content, Pronunciation, Fluency
Answer Short Questions	Content
Summarize Written Text	Content, Form, Grammar, Vocabulary
Write Essay	Content, Form, Development, Structure & Coherence, Grammar, General Linguistic Range, Vocabulary Range, Spelling
Summarize Spoken Text	Content, Form, Grammar, Vocabulary

What is Transcription?

- Orthographic representations (word-for-word) of student responses
- Annotated in detail for various speech events using specific rules

e.g.

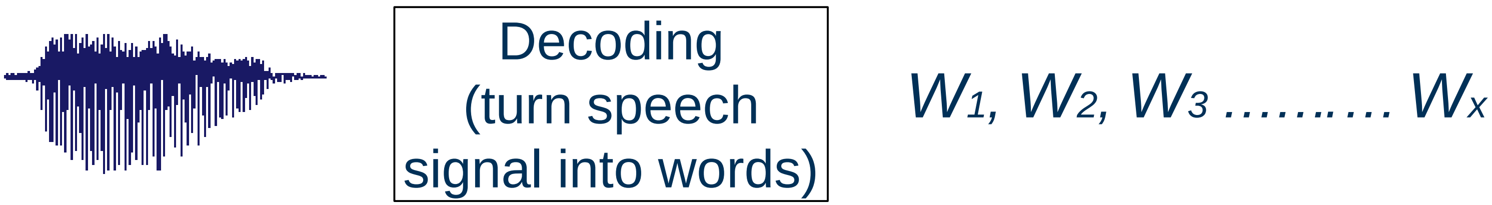
*[N] my *neighbor is trave(ling)- @ traveling this week*

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)**

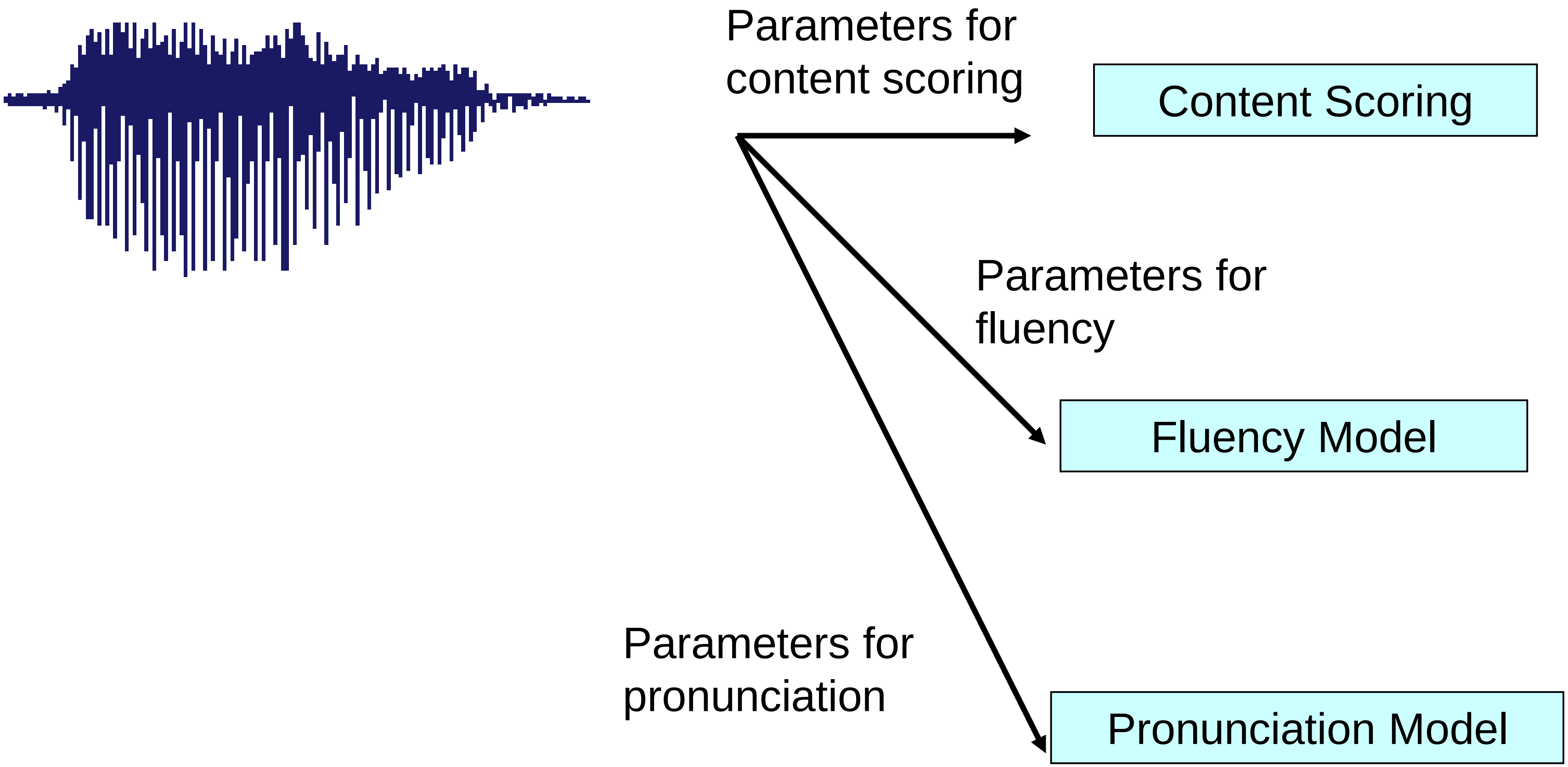


Decoding (turn speech signal into words)

W_1, W_2, W_3 ……… W_x

Automated Scoring – Speaking

Multiple Aspects per Response



Parameters for content scoring

Content Scoring

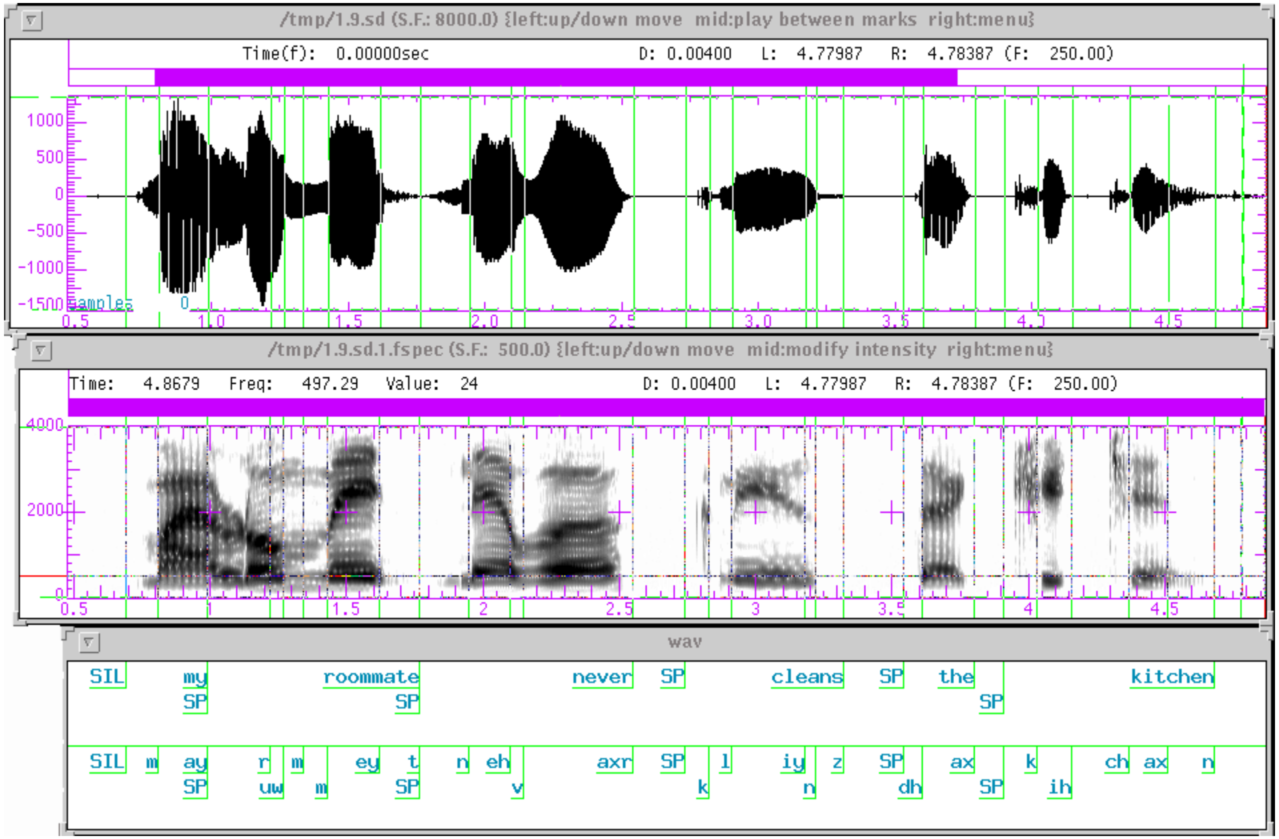
Parameters for fluency

Fluency Model

Parameters for pronunciation

Pronunciation Model

Automatic Speech Recognition



/tmp/1.9.sd (S.F.: 8000.0) {left:up/down move mid:play between marks right:menu}

Time(f): 0.00000sec D: 0.00400 L: 4.77987 R: 4.78387 (F: 250.00)

/tmp/1.9.sd.1.fspec (S.F.: 500.0) {left:up/down move mid:modify intensity right:menu}

Time: 4.8679 Freq: 497.29 Value: 24 D: 0.00400 L: 4.77987 R: 4.78387 (F: 250.00)

Waveform

Spectrum

Words
Segmentation

w1 w2 w3 w4 w5 w6

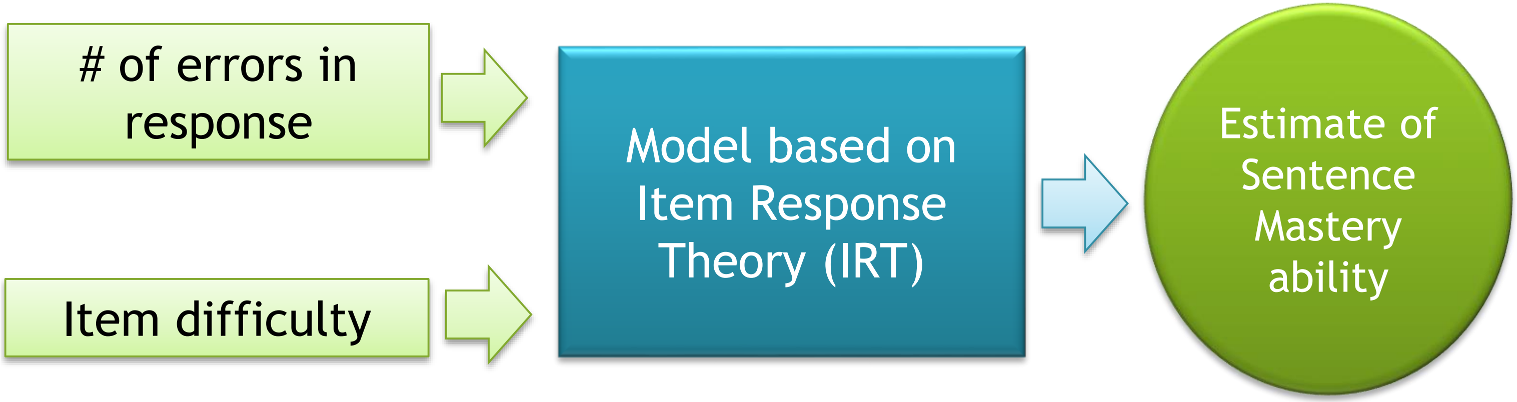
75-90 Words/Min

p p pppp p p p p pp ppp pp p p p p p

5.8 Phones/Sec

Content Scoring

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



# of errors in response

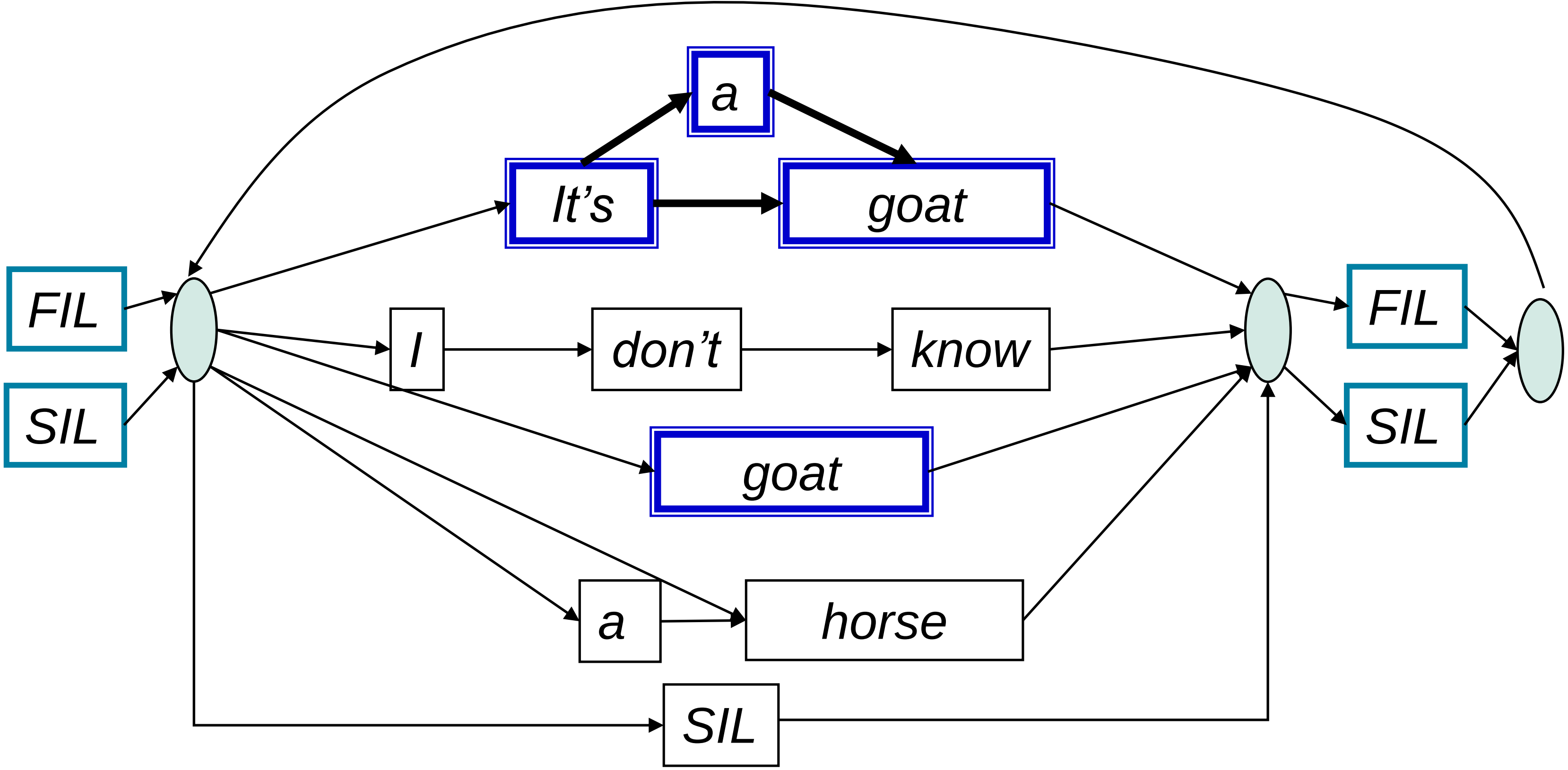
Item difficulty

Model based on Item Response Theory (IRT)

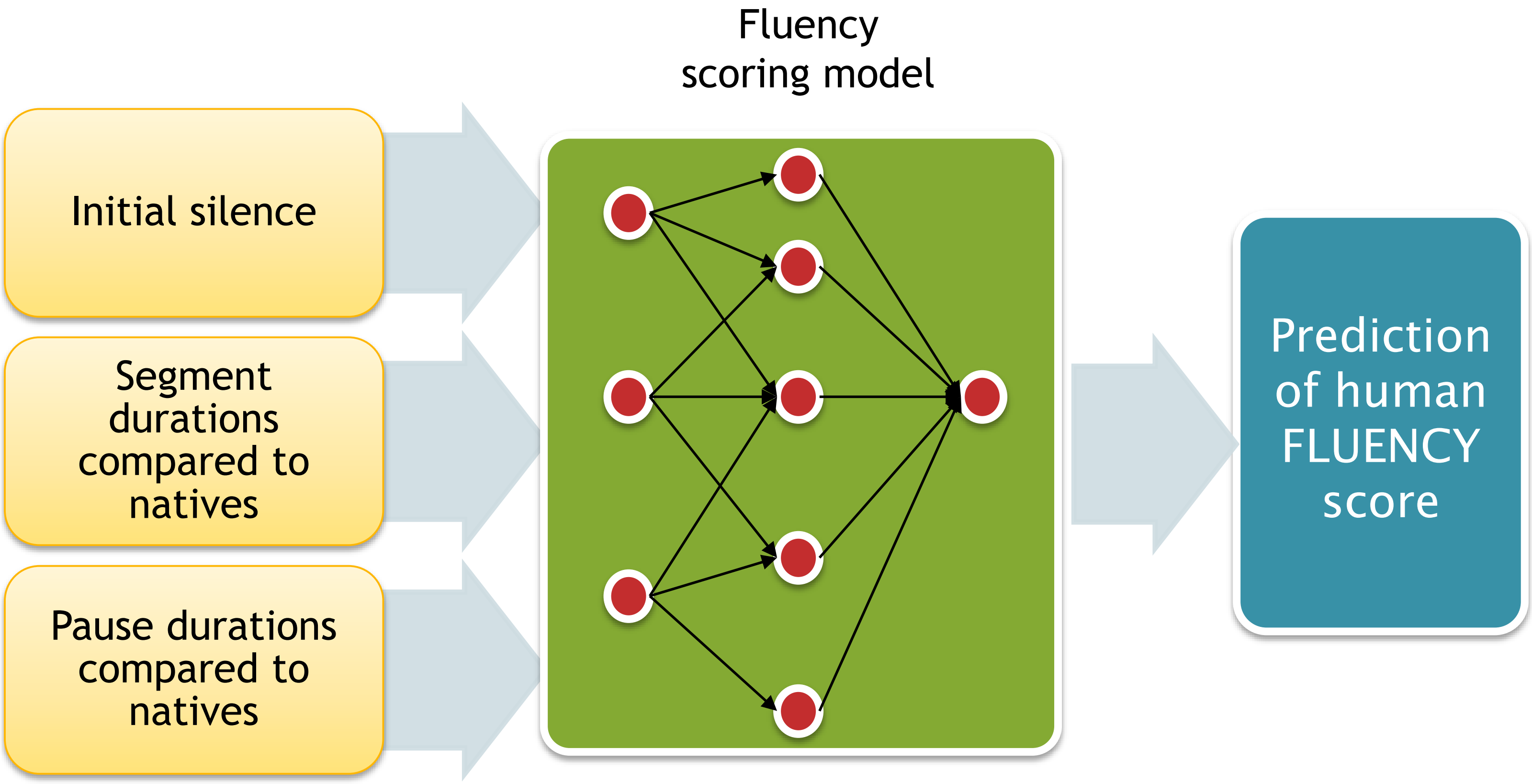
Estimate of Sentence Mastery ability

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

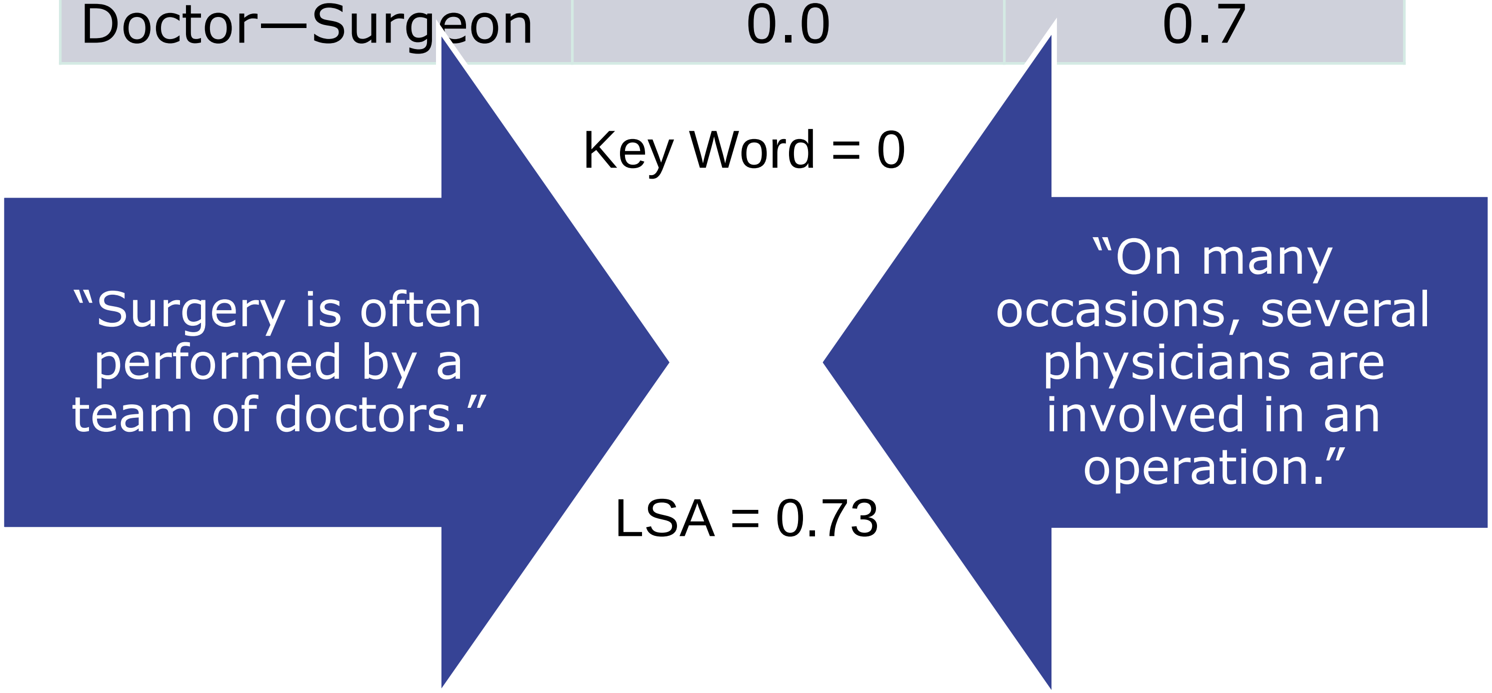
Prediction of human FLUENCY score

Automated Scoring – Writing

Pearson

Latent Semantic Analysis (LSA)

	Key Word	LSA
Doctor—Doctor	1.0	1.0
Doctor—Physician	0.0	0.8
Doctor—Surgeon	0.0	0.7



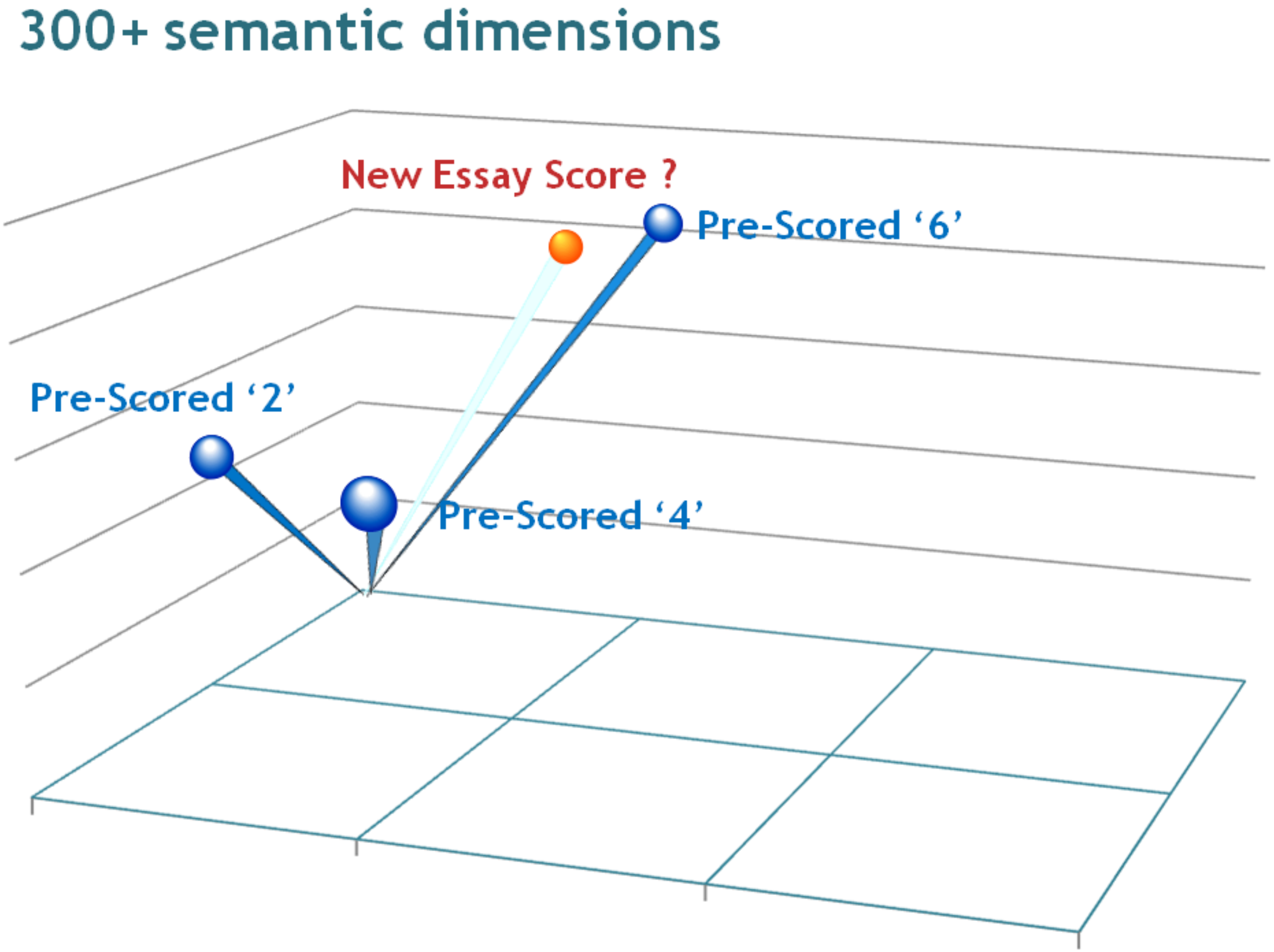
“Surgery is often performed by a team of doctors.”

Key Word = 0

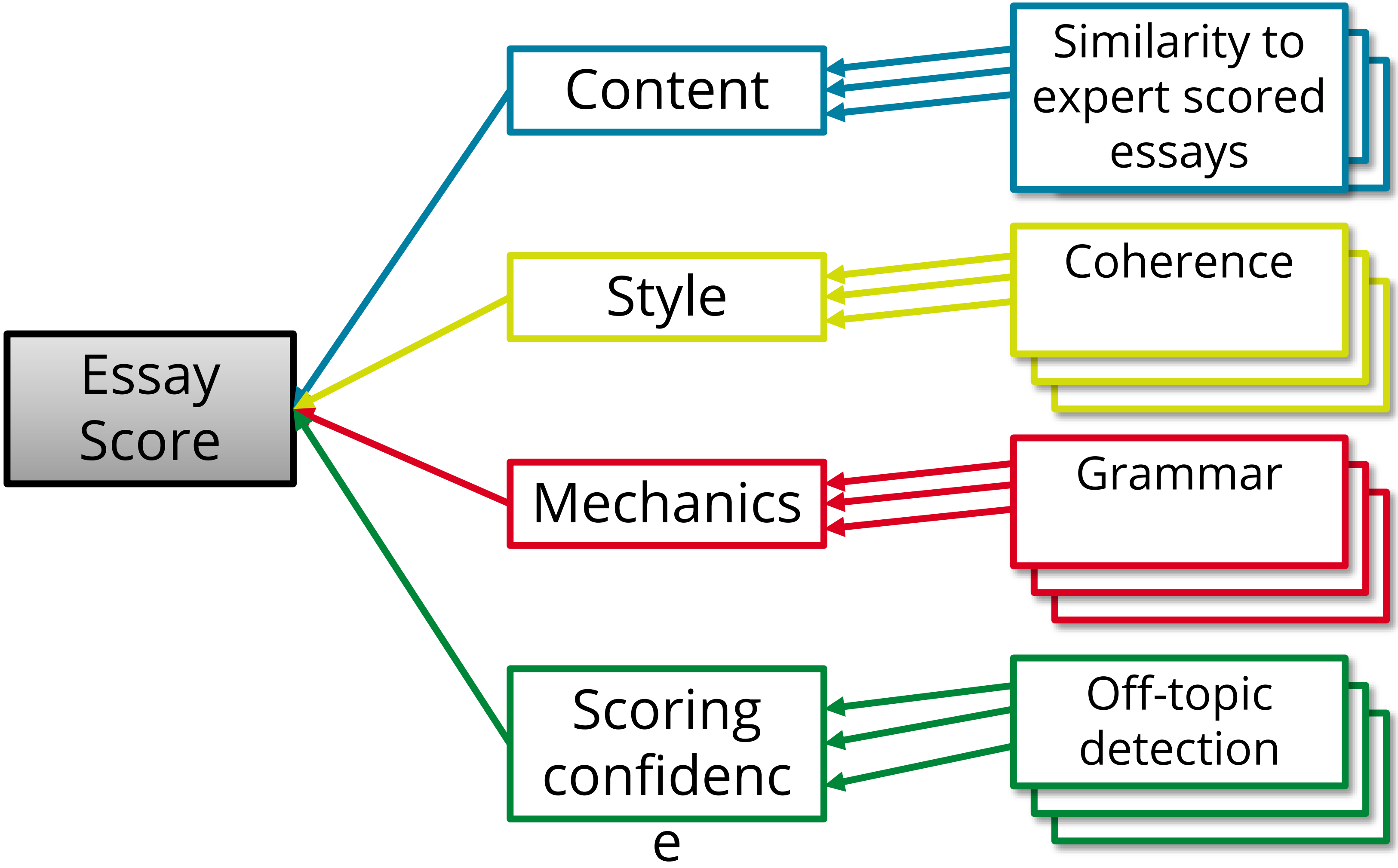
LSA = 0.73

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

Semantic Space for Essay Scoring



Essay scoring process



Essay Score

Content

Style

Mechanics

Scoring confidenc e

Similarity to expert scored essays

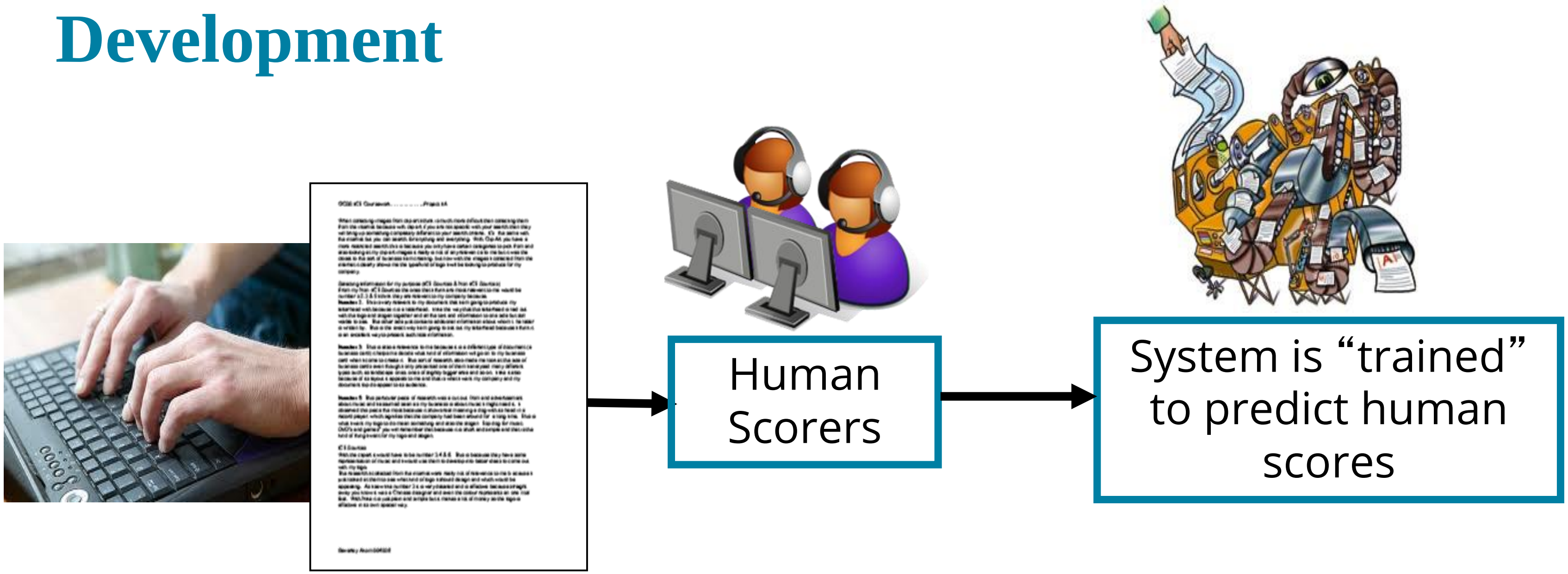
Coherence

Grammar

Off-topic detection

Validation

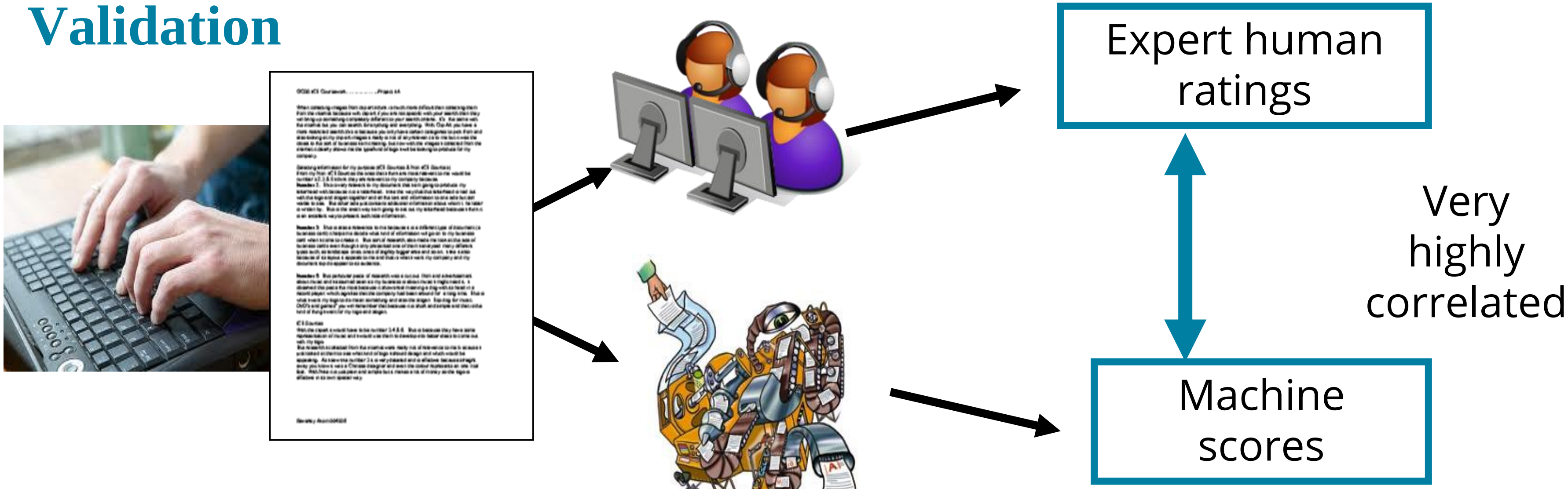
Development



Human Scorers

System is "trained" to predict human scores

Validation



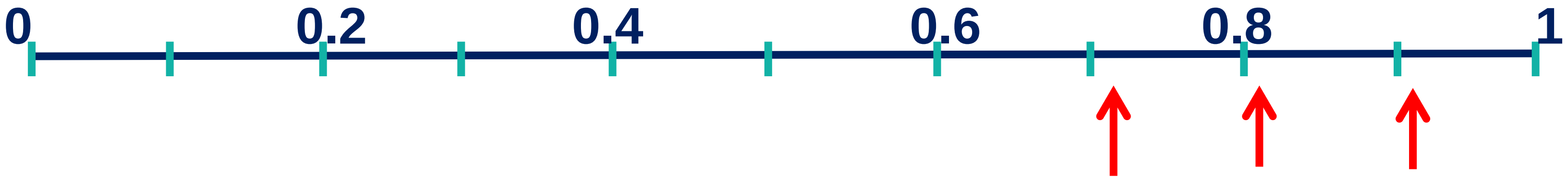
Expert human ratings

Very highly correlated

Machine scores

Test Reliability

	PTE Academic	IELTS	TOEFL
Overall	0.97	0.96	0.94
Reading	0.92	0.90	0.85
Listening	0.91	0.91	0.85
Writing	0.91	0.81-0.90	0.74
Speaking	0.91	0.83-0.86	0.88

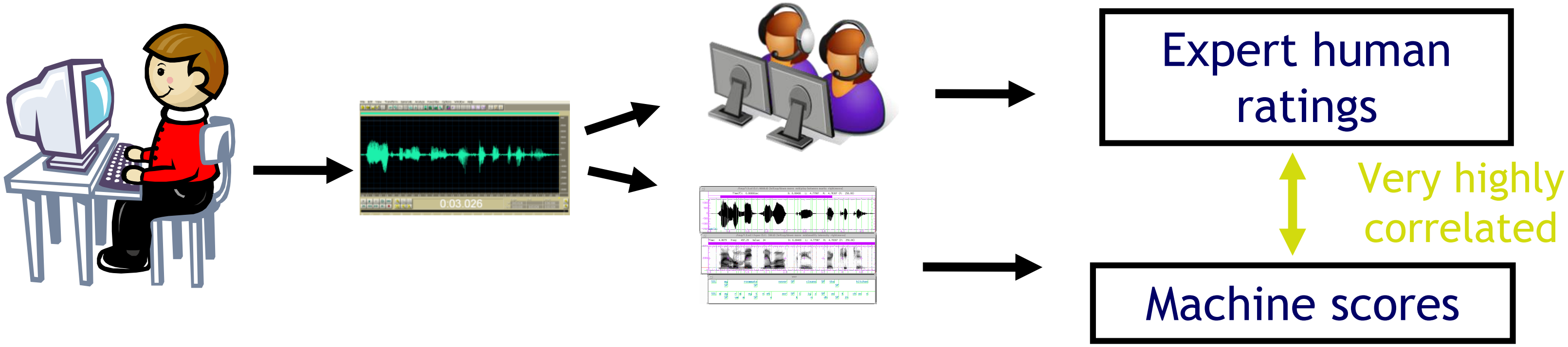


Acceptable

Good

Very Good

PTE-A Speaking Scores – Accuracy



Expert human ratings

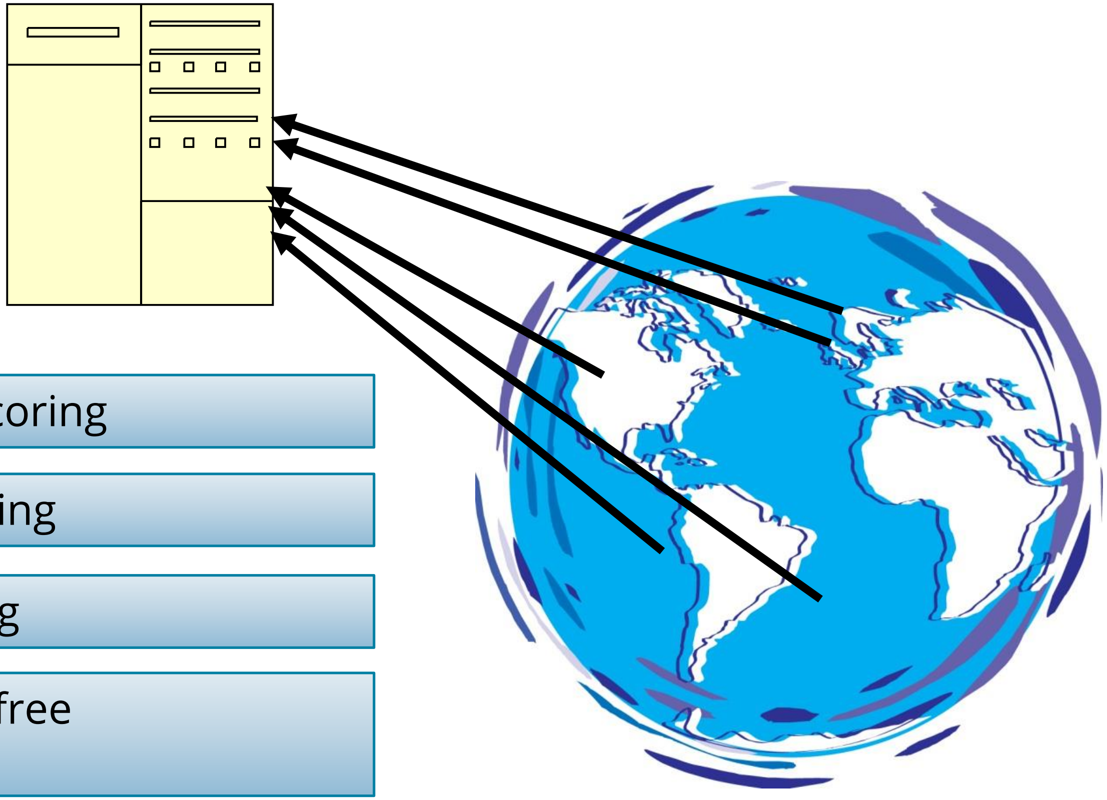
Very highly correlated

Machine scores

	Machine-Human Correlation (N=158)
Pronunciation	0.81
Fluency	0.82
Content	0.92
Vocabulary	0.90
Accuracy	0.95
Overall	0.96

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

VERSANT

tell™ test of english language learning

part of the aimsweb family

Progress

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

VERSANT™

Pearson

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?

November 6

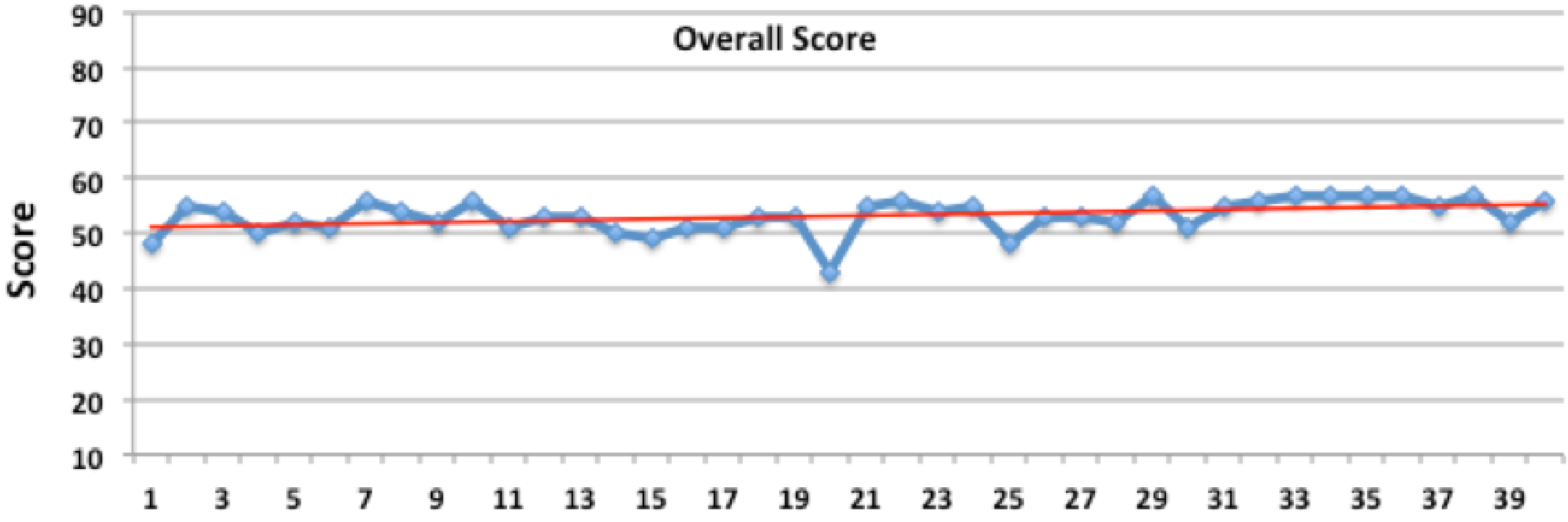
November 7

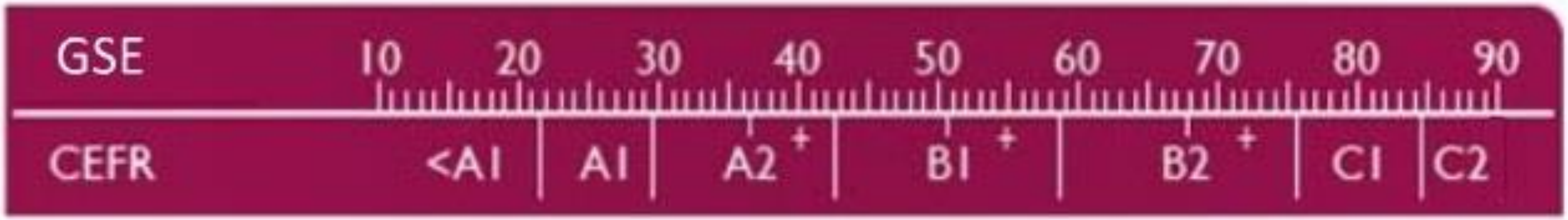
November 8

November 9

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.

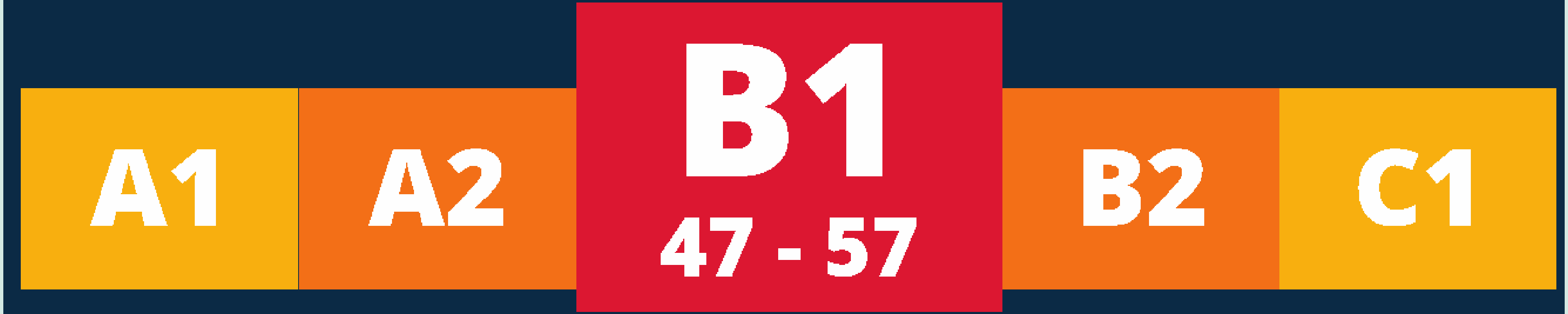
2. The final two tests (the ones with the highest scores) were not Mr. W's at all! Listening to the tests in the Versant database, it was apparent that their voices are different.

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.



A1

A2

B1

47 - 57

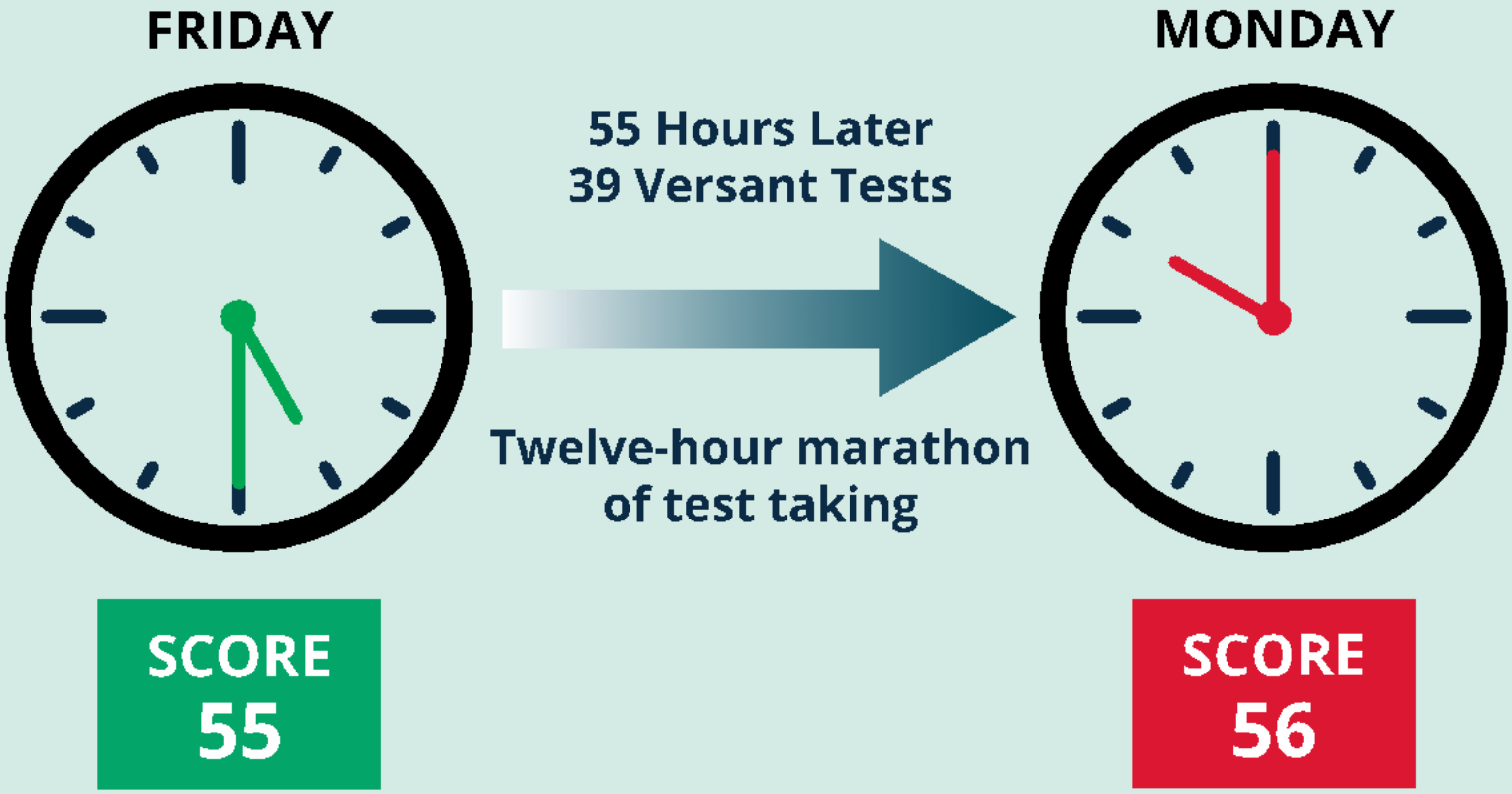
B2

C1

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

MONDAY

55 Hours Later
39 Versant Tests

Twelve-hour marathon
of test taking

SCORE
55

SCORE
56

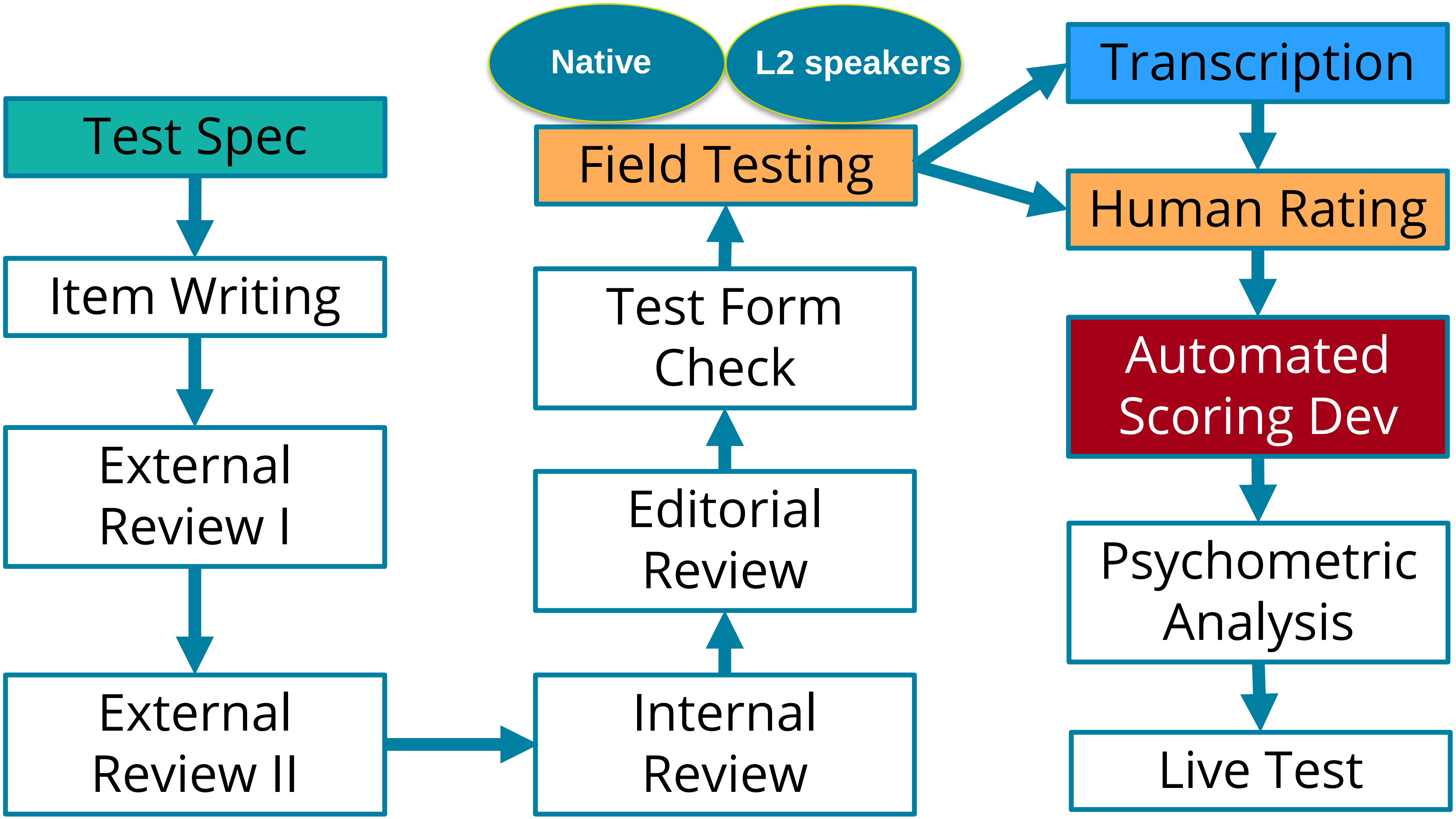
*"Despite its name, there is **nothing "artificial"** about this technology — it is made by humans, intended to behave like humans and affects humans. …. **No technology is more reflective of its creators than A.I**. It has been said that there are no "machine" values at all, in fact; machine values 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

Test Development Process



Test Spec

Item Writing

External Review I

External Review II

Internal Review

Editorial Review

Test Form Check

Field Testing

Native

L2 speakers

Transcription

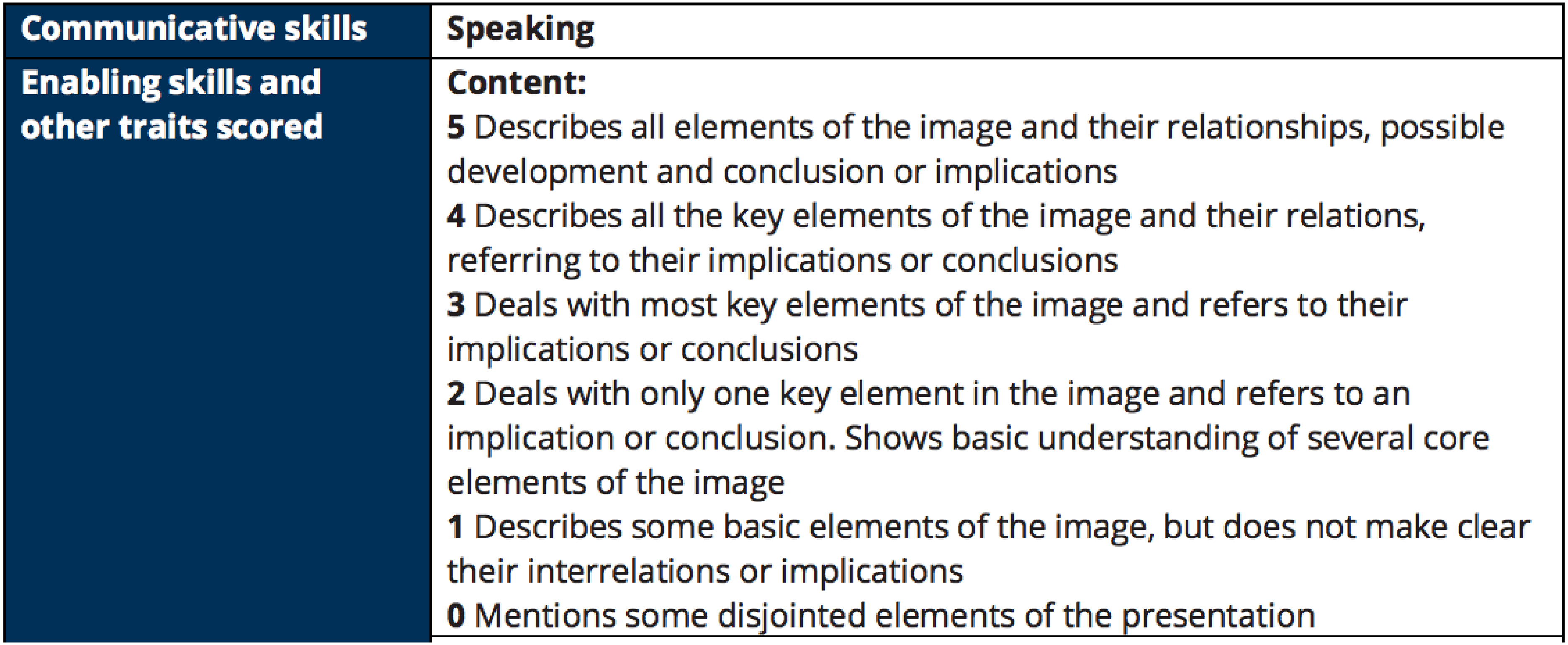
Human Rating

Automated Scoring Dev

Psychometric Analysis

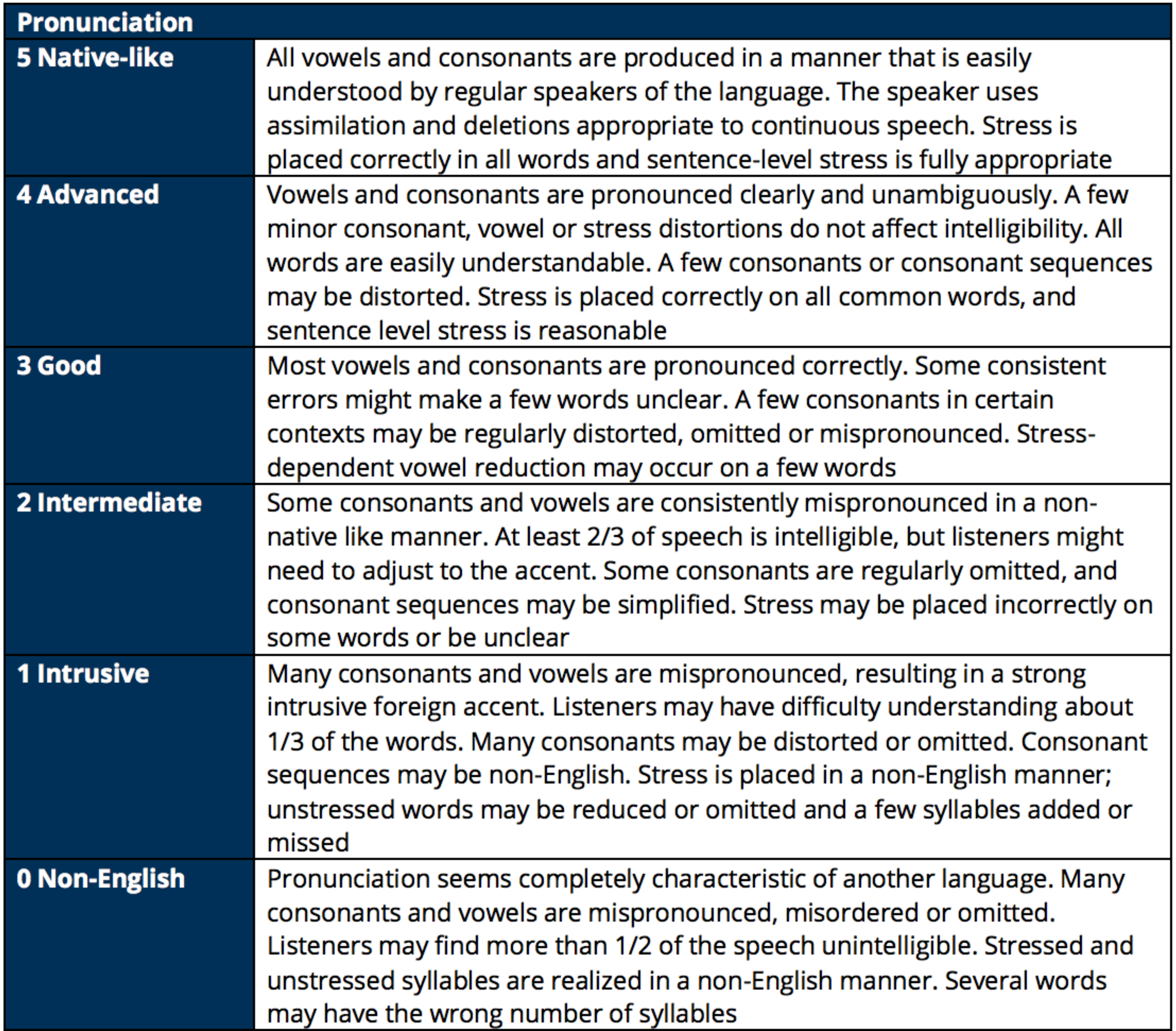
Live Test

Content Rubric



Communicative skills	Speaking
Enabling skills and other traits scored	**Content:** **5** Describes all elements of the image and their relationships, possible development and conclusion or implications **4** Describes all the key elements of the image and their relations, referring to their implications or conclusions **3** Deals with most key elements of the image and refers to their implications or conclusions **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 **1** Describes some basic elements of the image, but does not make clear their interrelations or implications **0** Mentions some disjointed elements of the presentation

Pronunciation



Pronunciation	
5 Native-like	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
4 Advanced	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
3 Good	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
2 Intermediate	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
1 Intrusive	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
0 Non-English	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

Content

2	The response provides a good summary of the text.
1	The response provides a fair summary of the text, but misses one or two aspects.
0	The response omits or misrepresents the main issue(s) dealt with in the text.
9	There is no response, response is not English or irrelevant

Development, Structure & Coherence

2	The essay shows a good development and logical structure
1	The essay is less well structured, some elements or paragraphs seem poorly linked
0	The essay lacks coherence, mainly consists of lists or loose elements
9	There is no response, response is not English or irrelevant

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