



Pearson

# 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?**









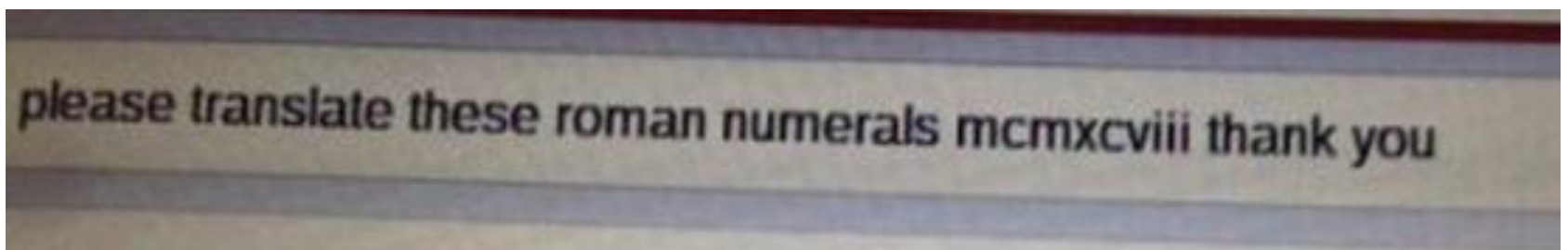
# Overview

1. General introduction of PTE Academic
1. Brief Overview of Test Development Process
2. Automated Scoring for Writing
3. Automated Scoring for Speaking





BEN JOHN/TWITTER  
@Push10Ben





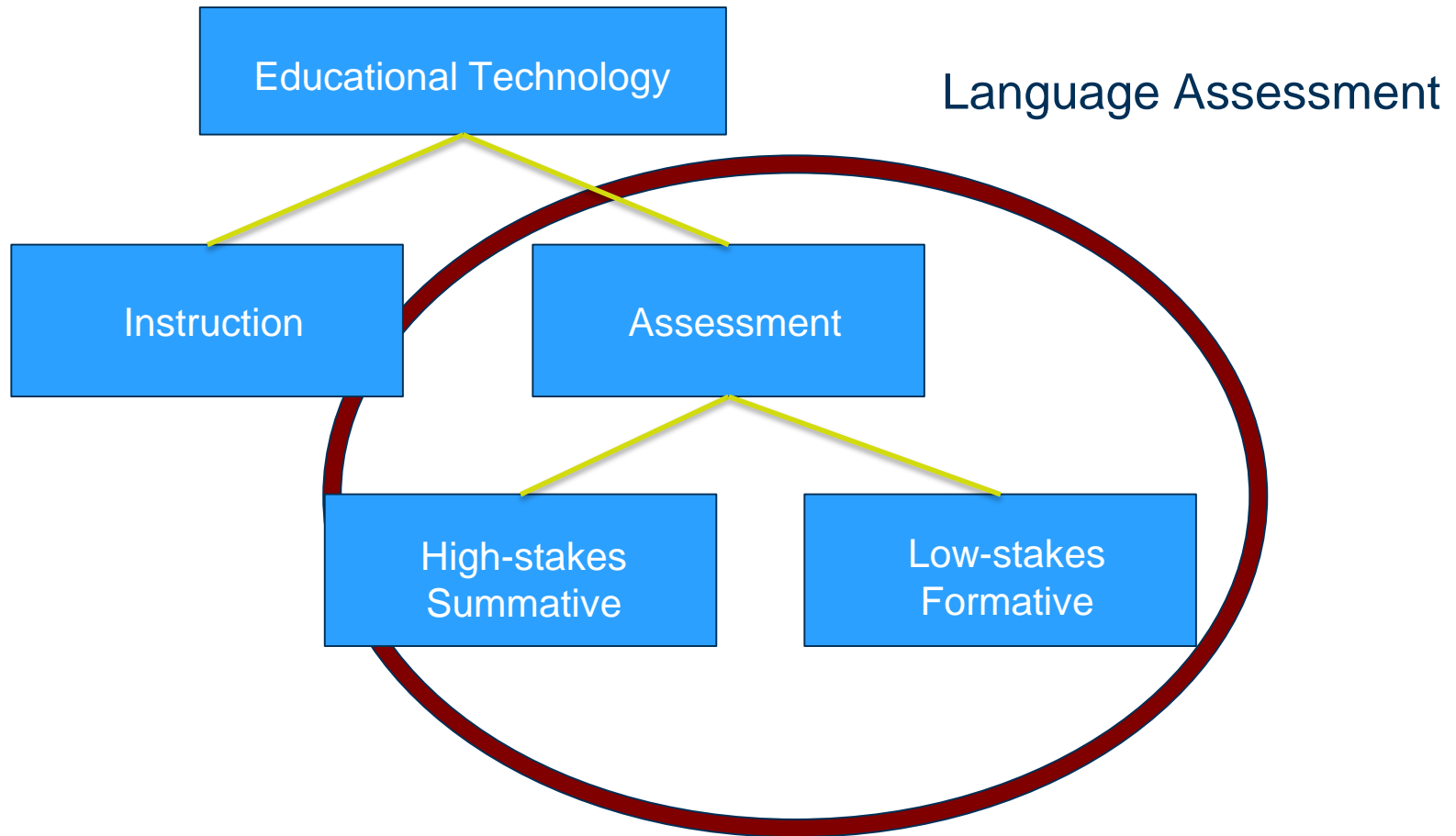


BEN JOHN/TWITTER  
@Push10Ben

**“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

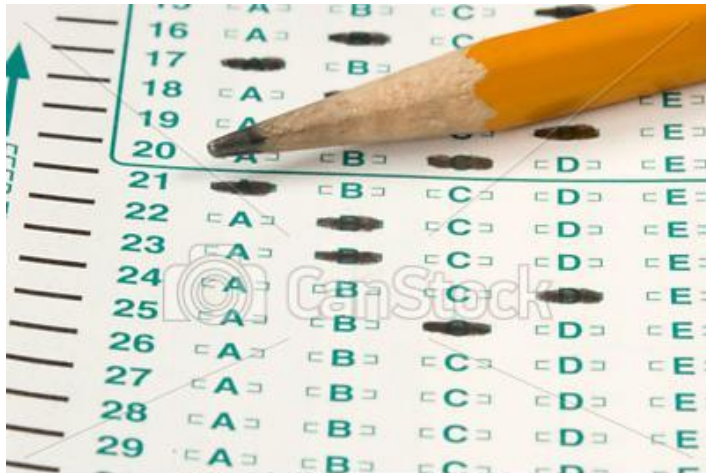




# 1970's – Technology!

## Bubble cards & readers

- Multi-choice format
- Passive skills



© Can Stock Photo – csp2386918



- 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



# PTE-A Score Report

**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.

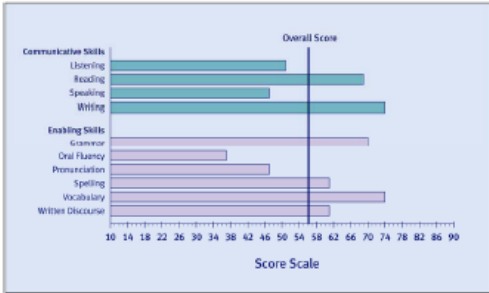
The graph below indicates this test taker's Communicative Skills and Enabling Skills relative to his or her Overall Score.

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](http://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](http://www.pearsonvue.com/ptescores)

© Elsevier 2009  
All rights reserved, no part of this publication may be reproduced without the prior written permission of Elsevier (a part of Pearson company)

## Overall score

## Communicative Skills

- Speaking
- Writing
- Reading
- Listening

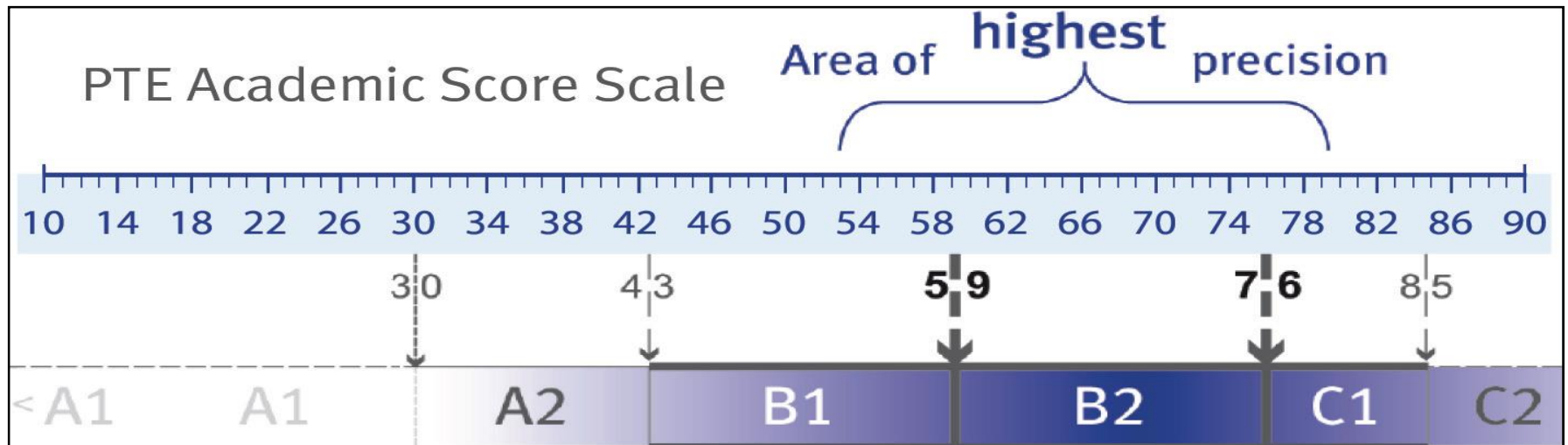
## Enabling Skills

- 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

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



**88%**  
of Canadian  
universities

**100%**  
of Irish  
universities



**96%**  
of UK  
universities



Accepted by  
universities in Asia

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

**100%**

of Australian universities,  
most professional associations and for all  
student or migration visas



**100%**

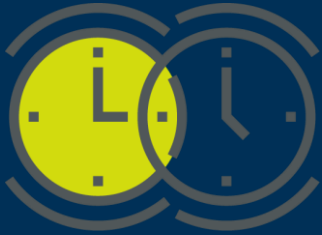
of New Zealand  
universities and for  
all student or migration visas



For a complete listing: [pearsonpte.com/accepts](https://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

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



# 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
<b>Summarize spoken text</b>	After listening to a recording, write a 50-70 word summary	listening and writing	60-90 seconds
<b>Multiple choice, choose multiple answers</b>	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
<b>Fill in the blanks</b>	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
<b>Highlight correct summary</b>	After listening to a recording, select the paragraph that best summarizes the recording	listening and reading	30-90 seconds
<b>Multiple choice, choose single answer</b>	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
<b>Select missing word</b>	After listening to a recording, select the missing word that completes the recording from a list of options	listening	20-70 seconds
<b>Highlight incorrect words</b>	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
<b>Write from dictation</b>	After listening to a recording of a sentence, type the sentence	listening and writing	3-5 seconds



# 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.



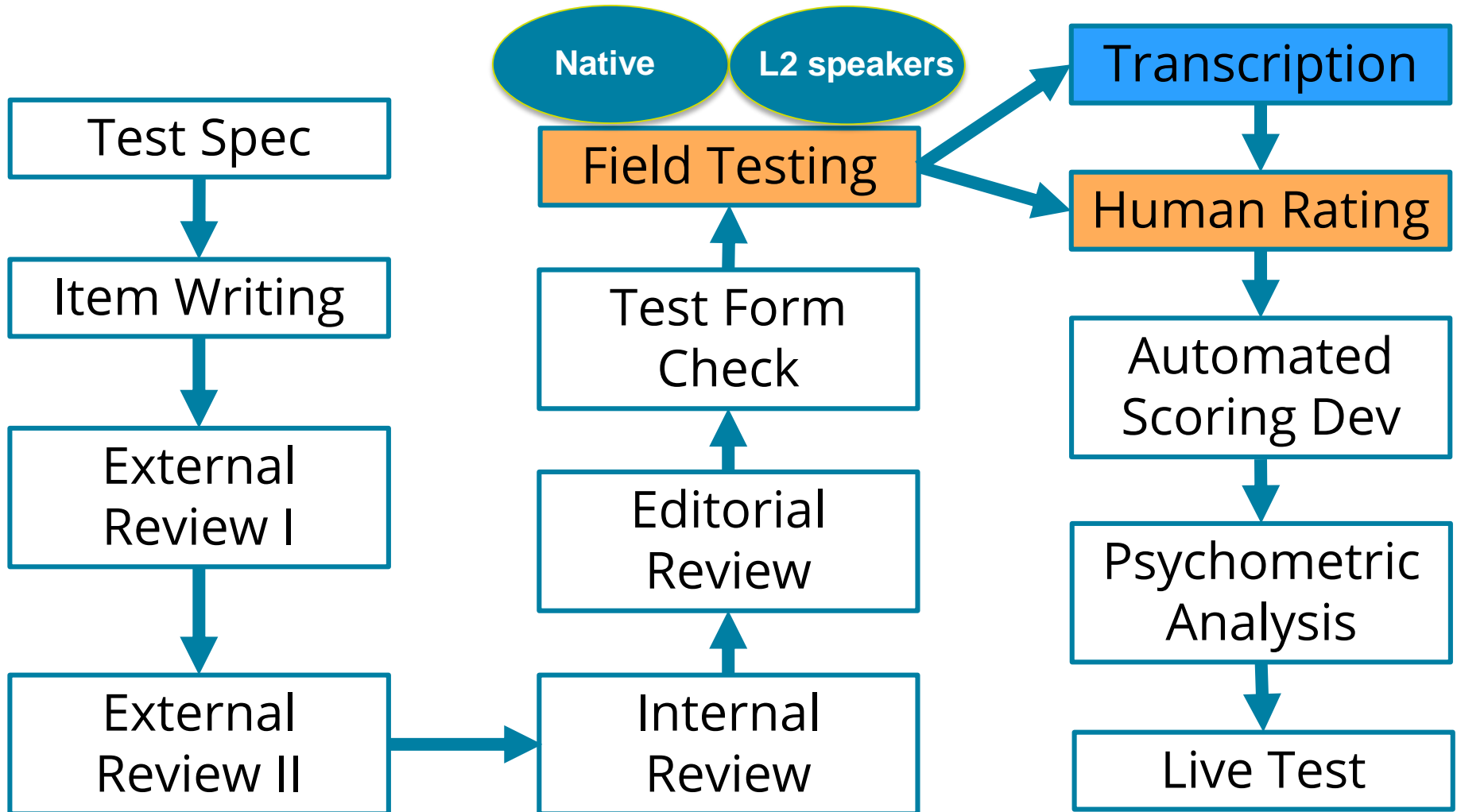
Including international and non-native English making the test more appropriate test for modern global universities.



**How did we develop the test?**



# Test Development Process





# 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

Automated Test	Correlation to human raters
Spanish	.97
Dutch	.93
Arabic	.98
French	.97
Chinese	.96
English	.97

## 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**

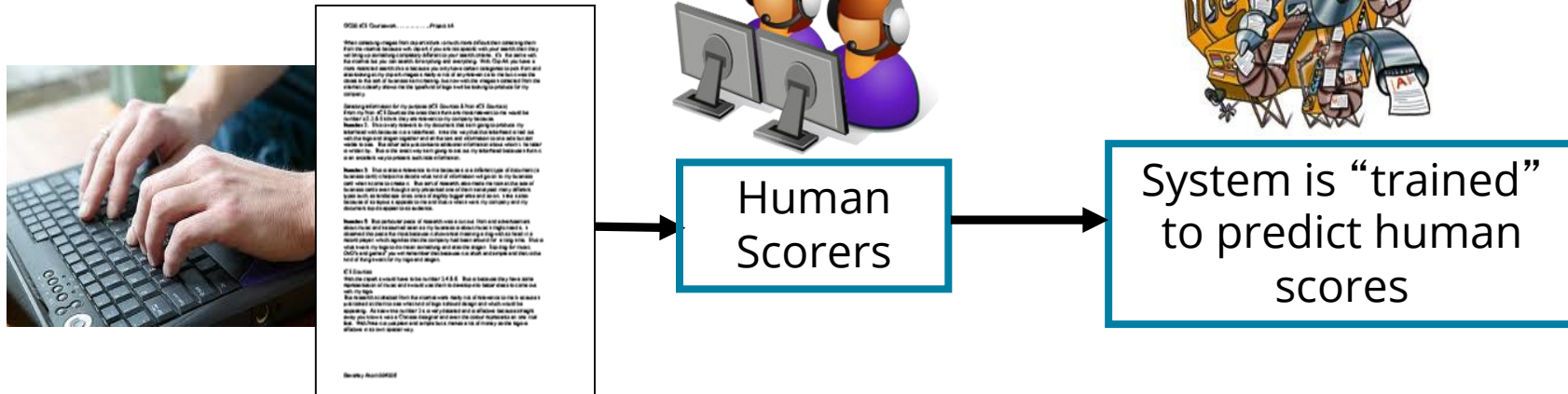


# Why automated scoring?

- 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





# 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)

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

Key Word = 0

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

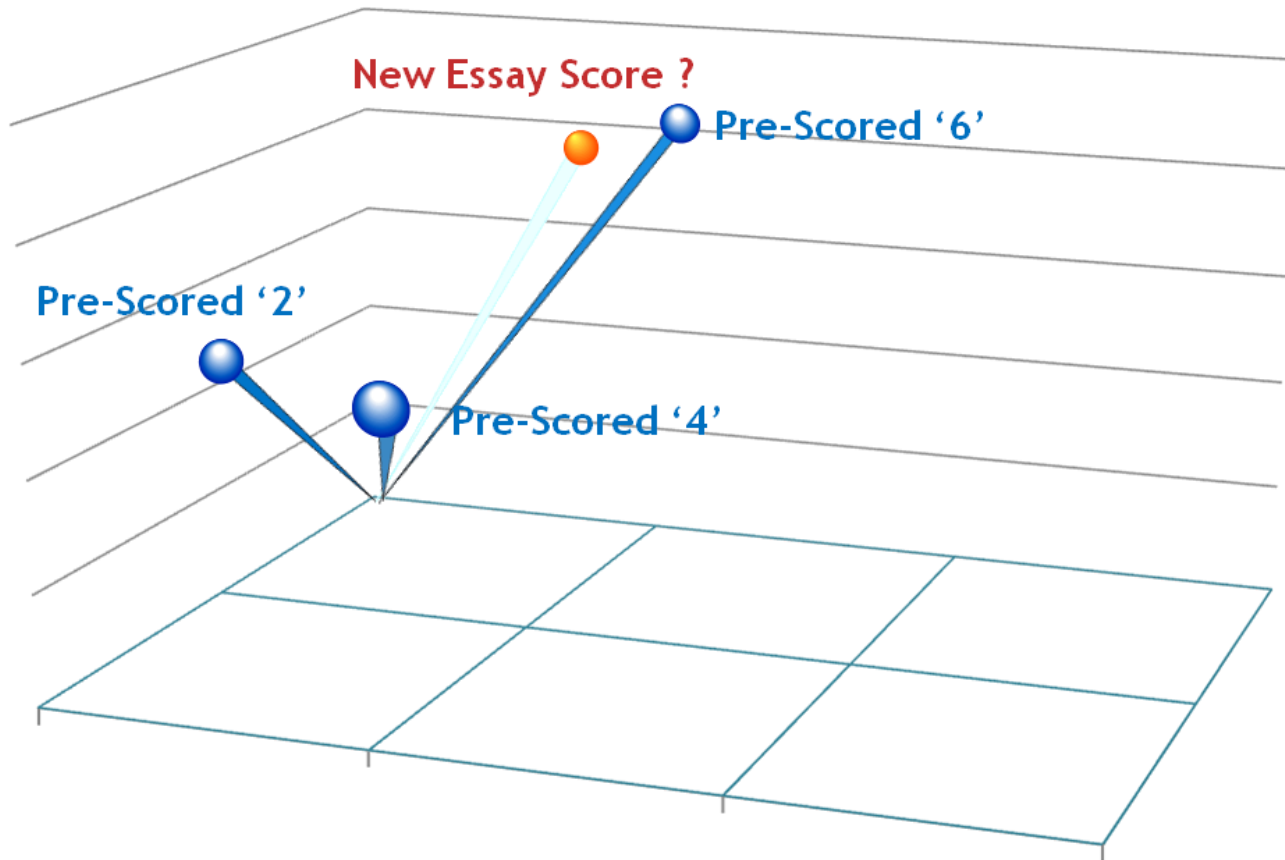
LSA = 0.73

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



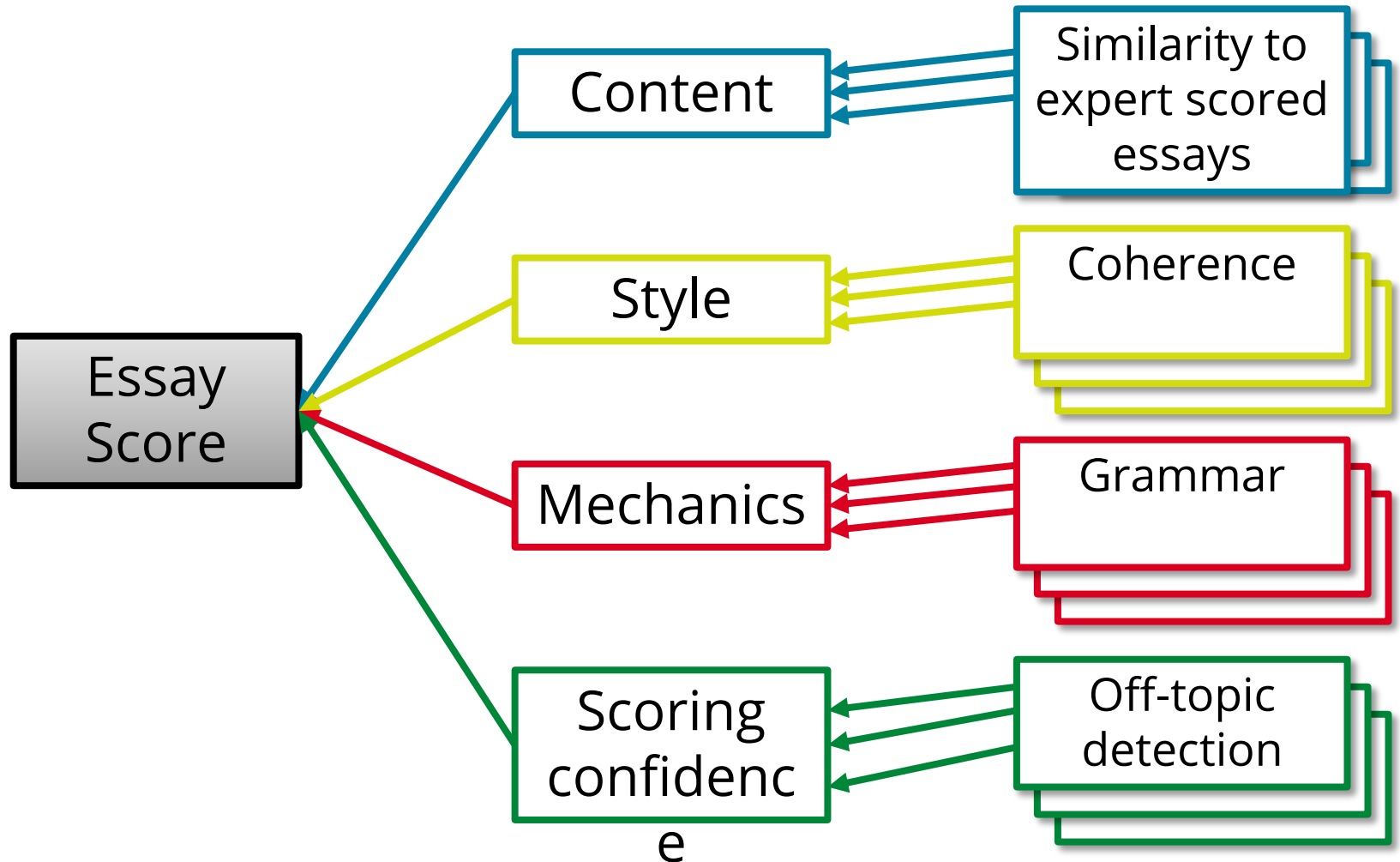
# Essay Scoring

300+ semantic dimensions





# Essay scoring process





# 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?

Cut

Copy

Paste

Total Word Count: 0



# 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



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 **temrs** 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 **wheter** they want to continue their life as a smoker and take all the risk, **or stop and learn to life healthier.**



# Score Comparison

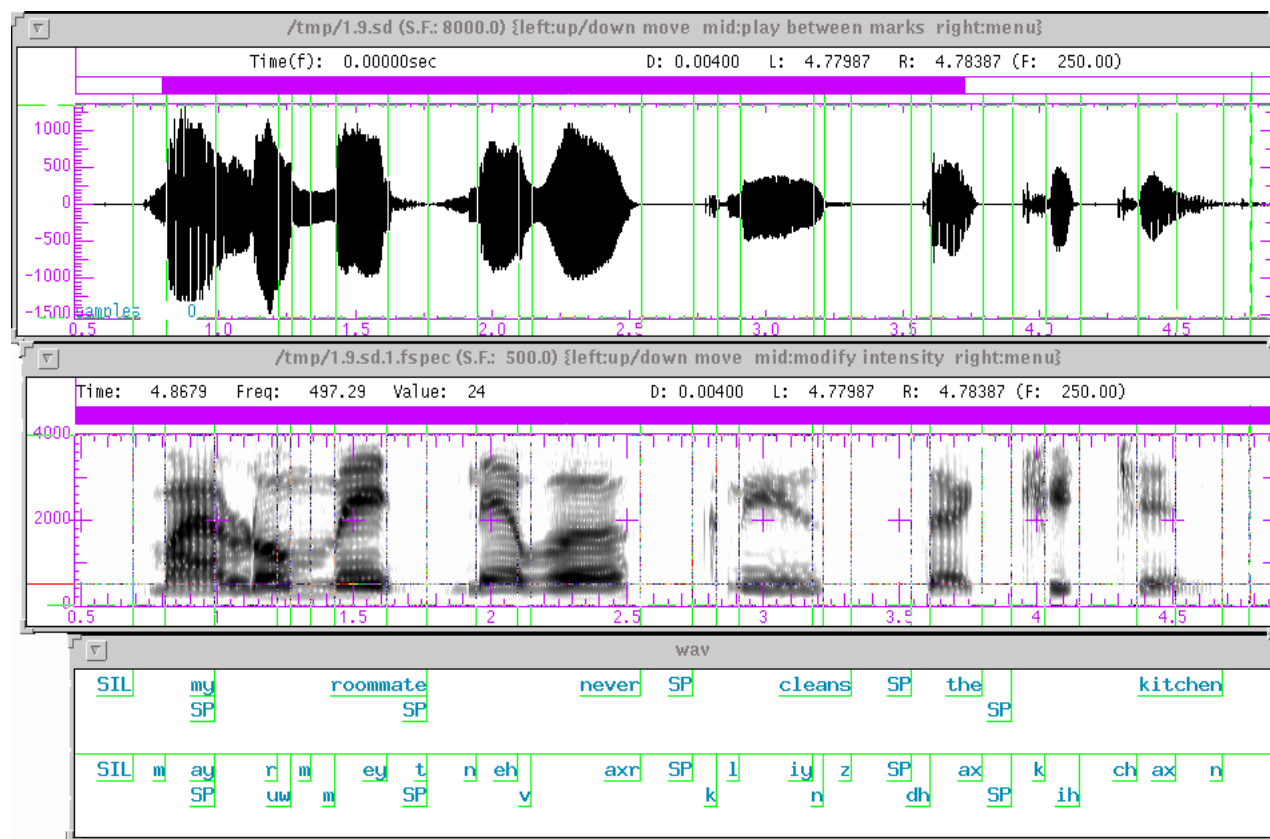
Trait	Human 1	Human 2	Adjudicator	IEA
Content	2	2		1.80
Development, Structure & Coherence	0	1	1	1.35
General Linguistic Range	1	1		1.03
Grammar Usage & Mechanics	1	1		1.07
Vocabulary Range	1	2	1	0.93



# **Automated Scoring – Speaking**



# Automatic Speech Recognition



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**

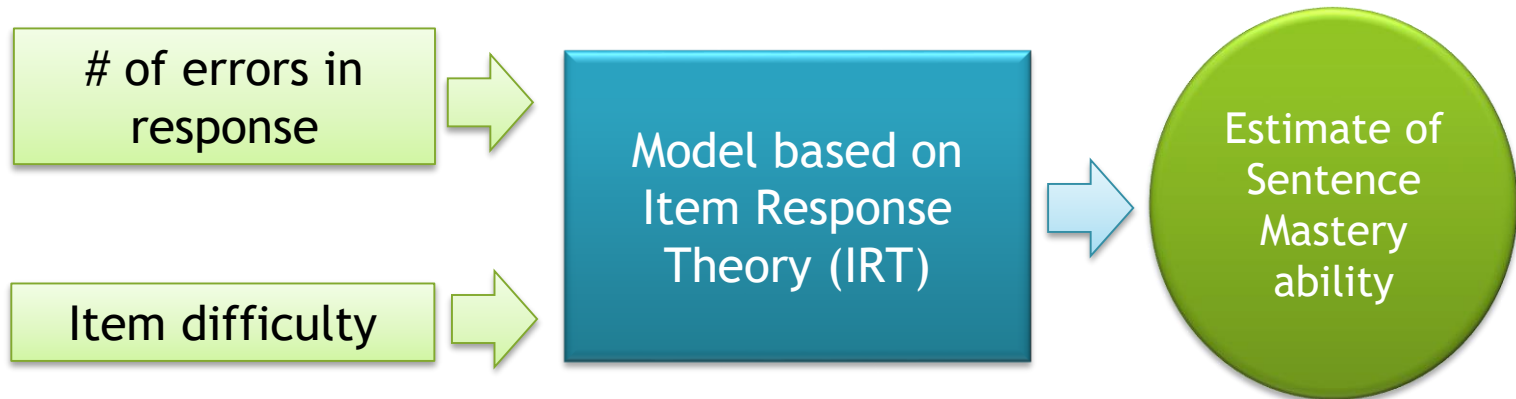
**5.8 Phones/Sec**



# Content Scoring

Correct Answer:

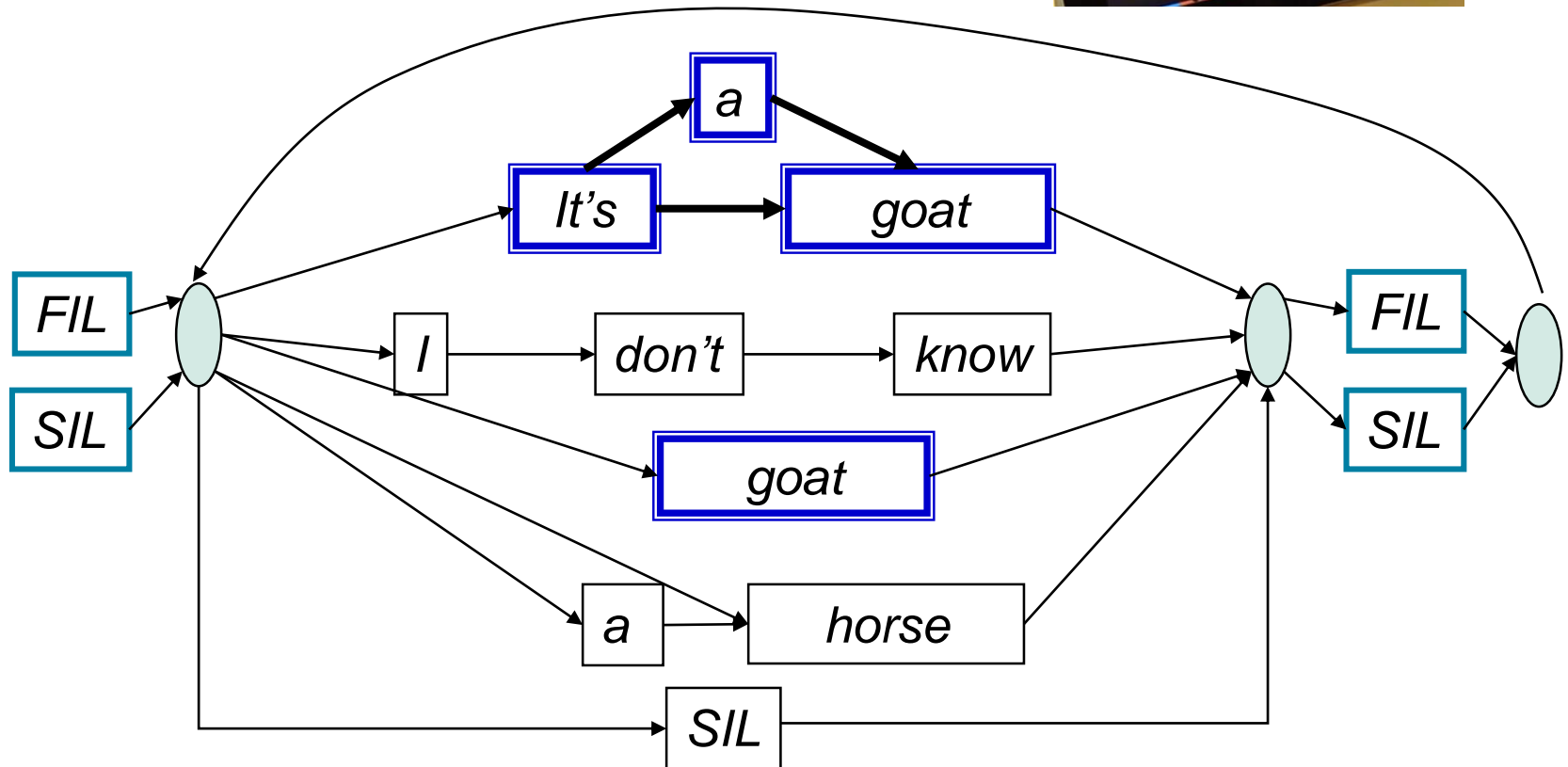
*“it’s supposed to rain tomorrow isn’t it”*





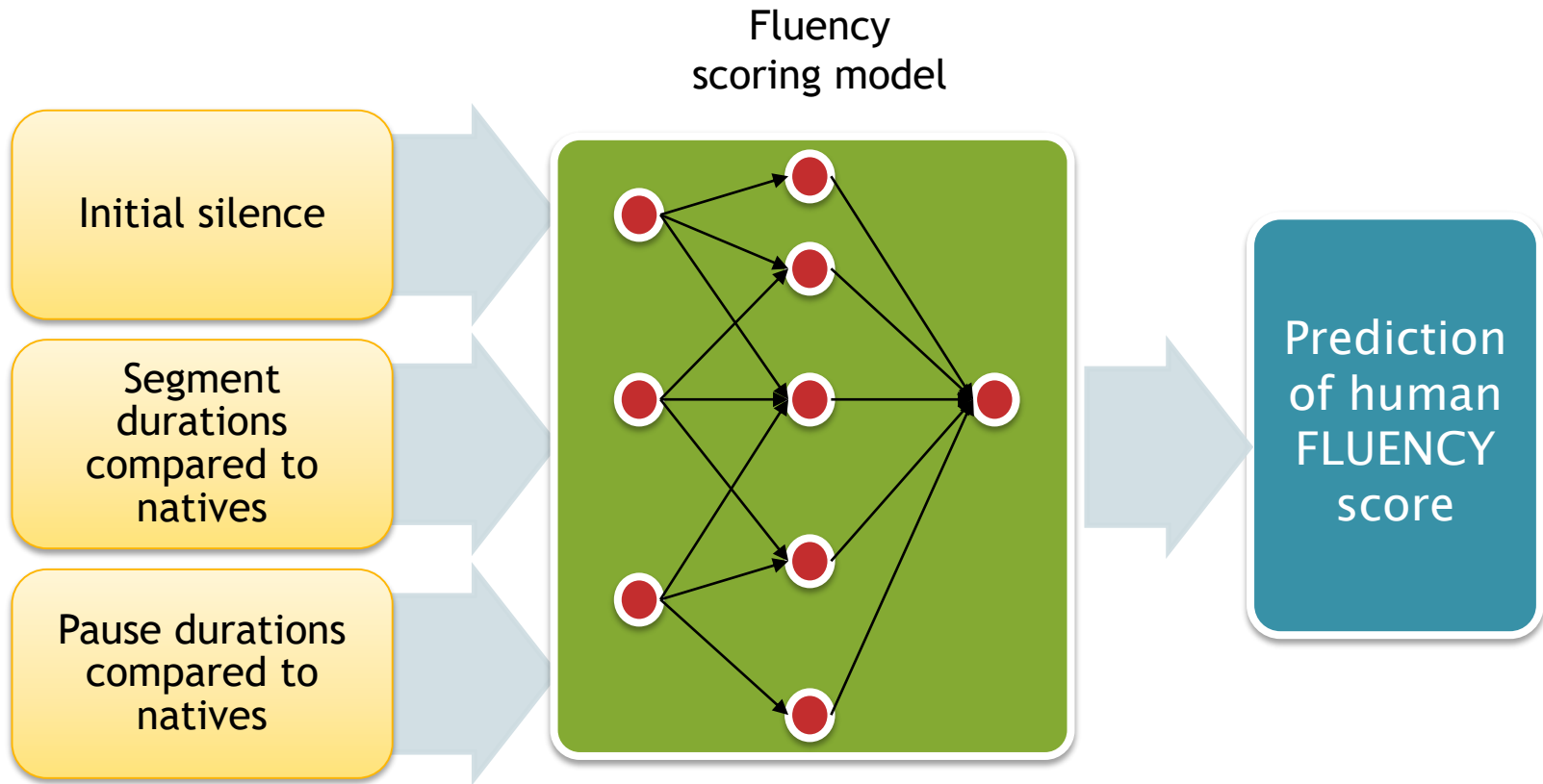
# Simplified Response Network

Example: *Say what's in the picture.*





# Manner Scoring

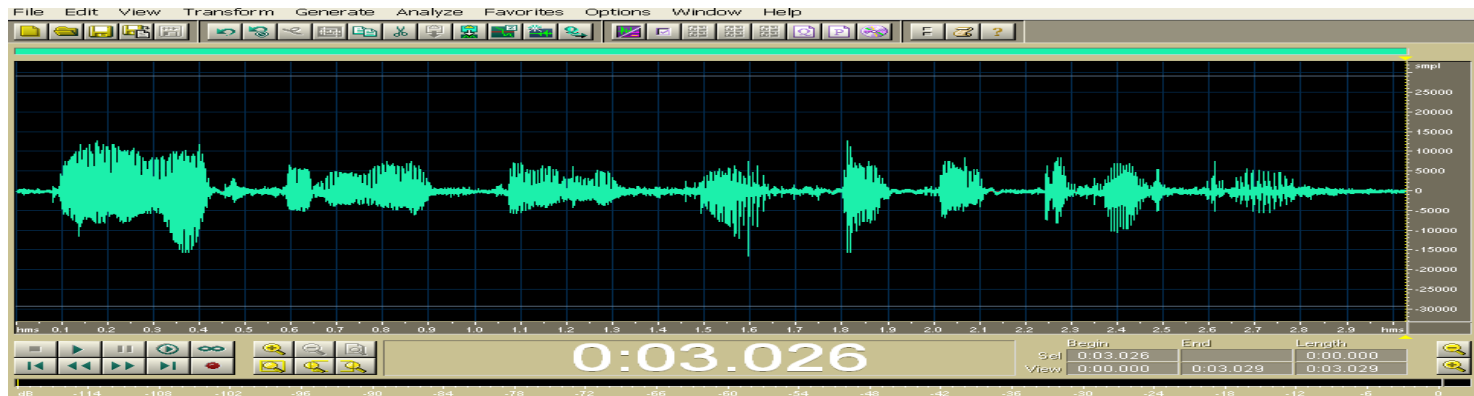






## Example: Native Speaker

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



*"New York city is famous for its ethnic diversity"*

<b>Pronunciation:</b>	<b>8.7</b>
<b>Fluency:</b>	<b>8.1</b>
<b>Accuracy:</b>	<b>0 word errors</b>





## Example: English Learner

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

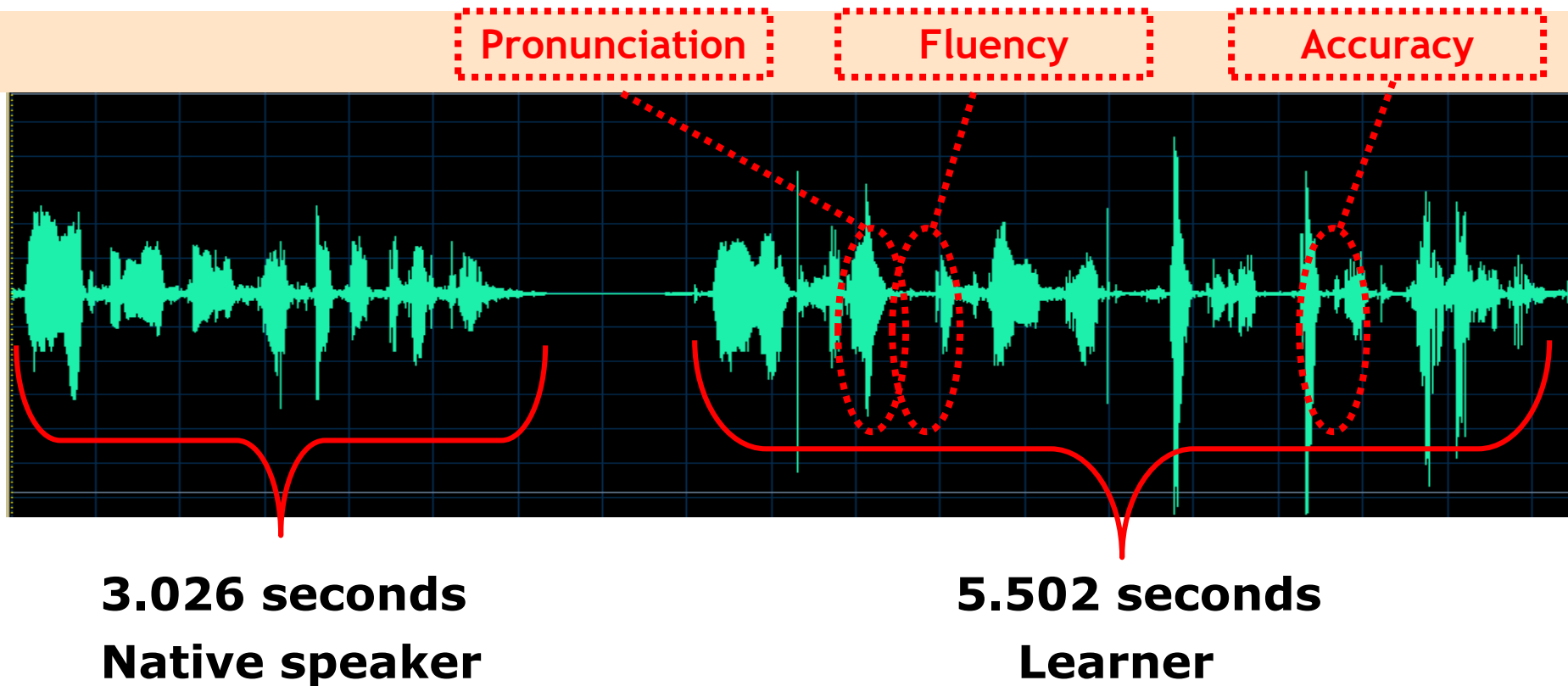


*"New York city is famous for its ethnic ethnic diversity"*

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

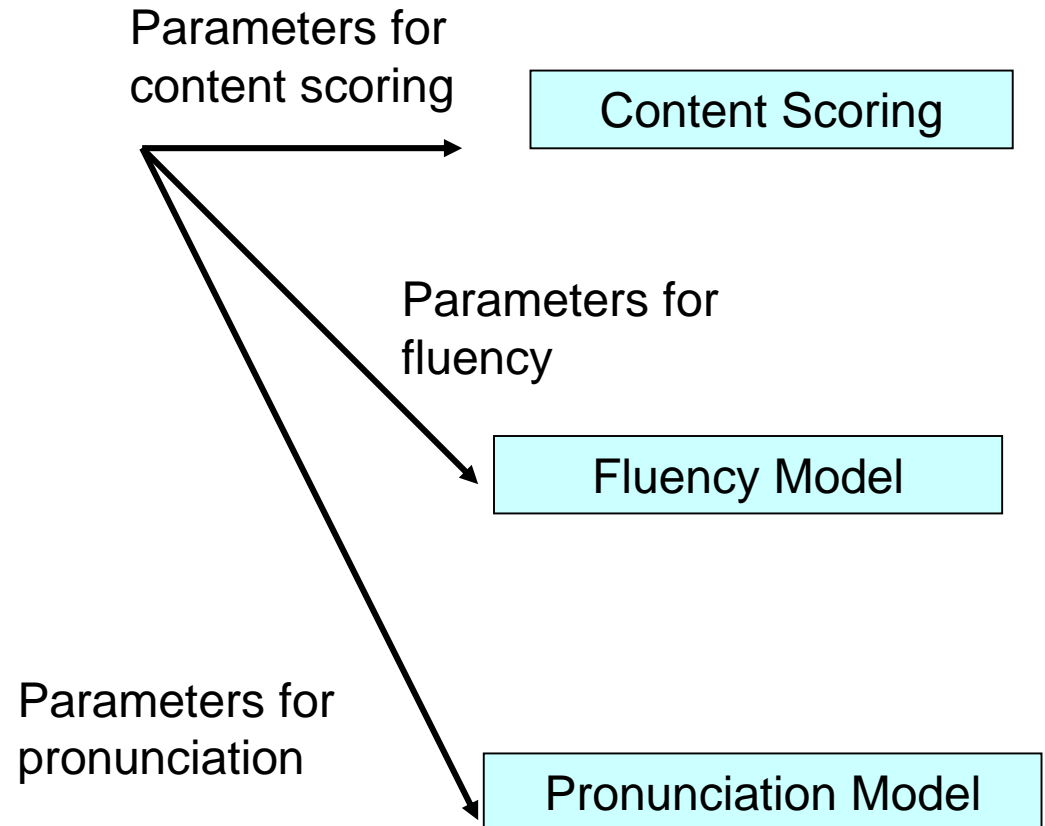
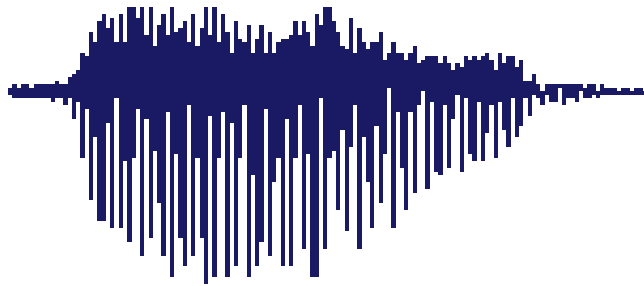


# Performance Comparison





# Multiple Aspects per Response





# Validation



# Development



Human  
Scorers



System is “trained”  
to predict human  
scores

# Validation



Expert human  
ratings

Very  
highly  
correlated

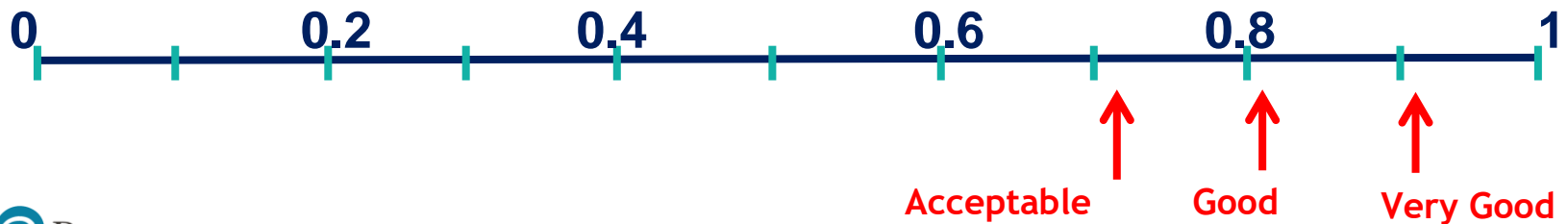


Machine  
scores



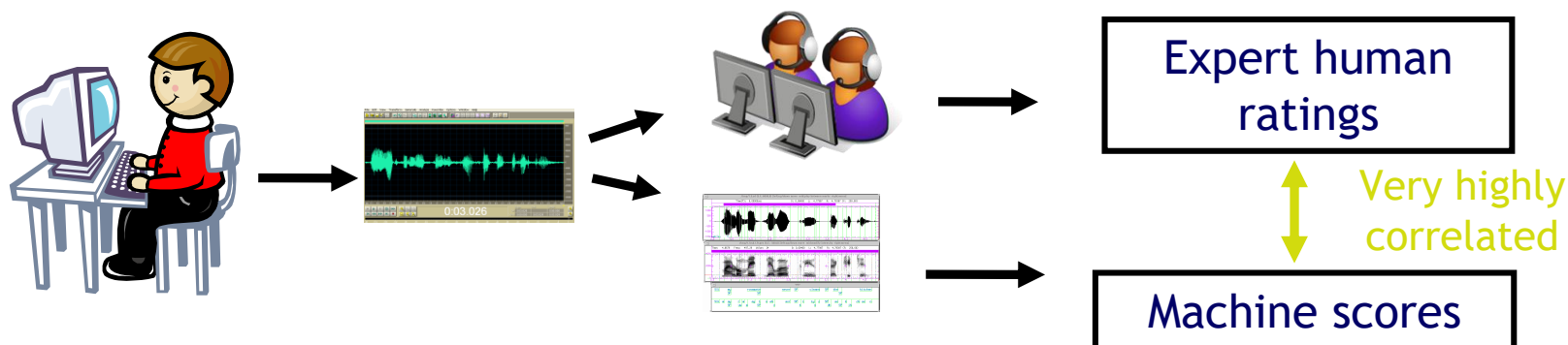
# Test Reliability

	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





# PTE-A Speaking Scores – Accuracy

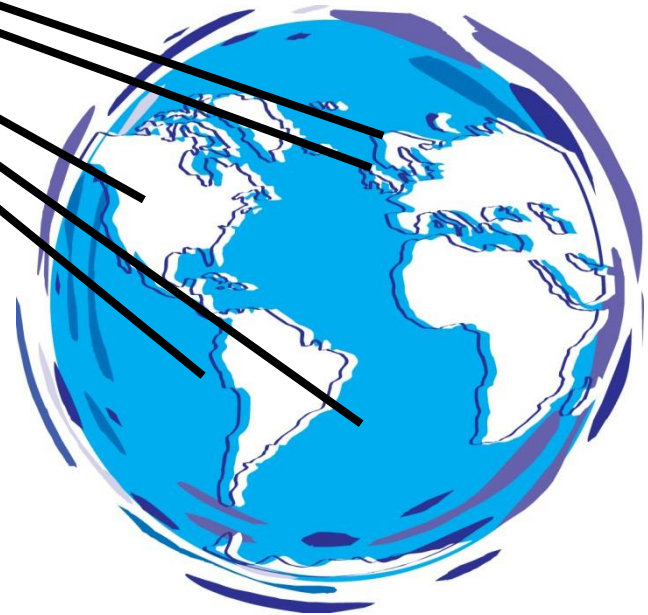
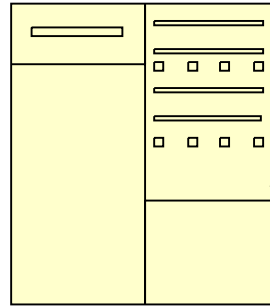


	Machine-Human Correlation (N=158)
Pronunciation	0.81
Fluency	0.82
Content	0.92
Vocabulary	0.90
Accuracy	0.95
Overall	0.96



# 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.

Book up to  
24 hours before.

Fast - 85% of results  
within 2 days.

## 2. Secure

Advanced multi  
layer security  
measures.

Most secure  
tests.

Confidence  
in results.

## 3. Accurate

Computer based  
marking ensures  
impartial,  
accurate marks.

Institutions can trust  
the English ability  
of students.

## 4. Relevant & Objective

Uses genuine  
academic content and  
integrated tasks testing  
multiple skills.

Reflects the use  
of language  
students need.



ALWAYS LEARNING