Automated scoring in large scale international testing

John de Jong
Language Testing Services
Velp, Netherlands

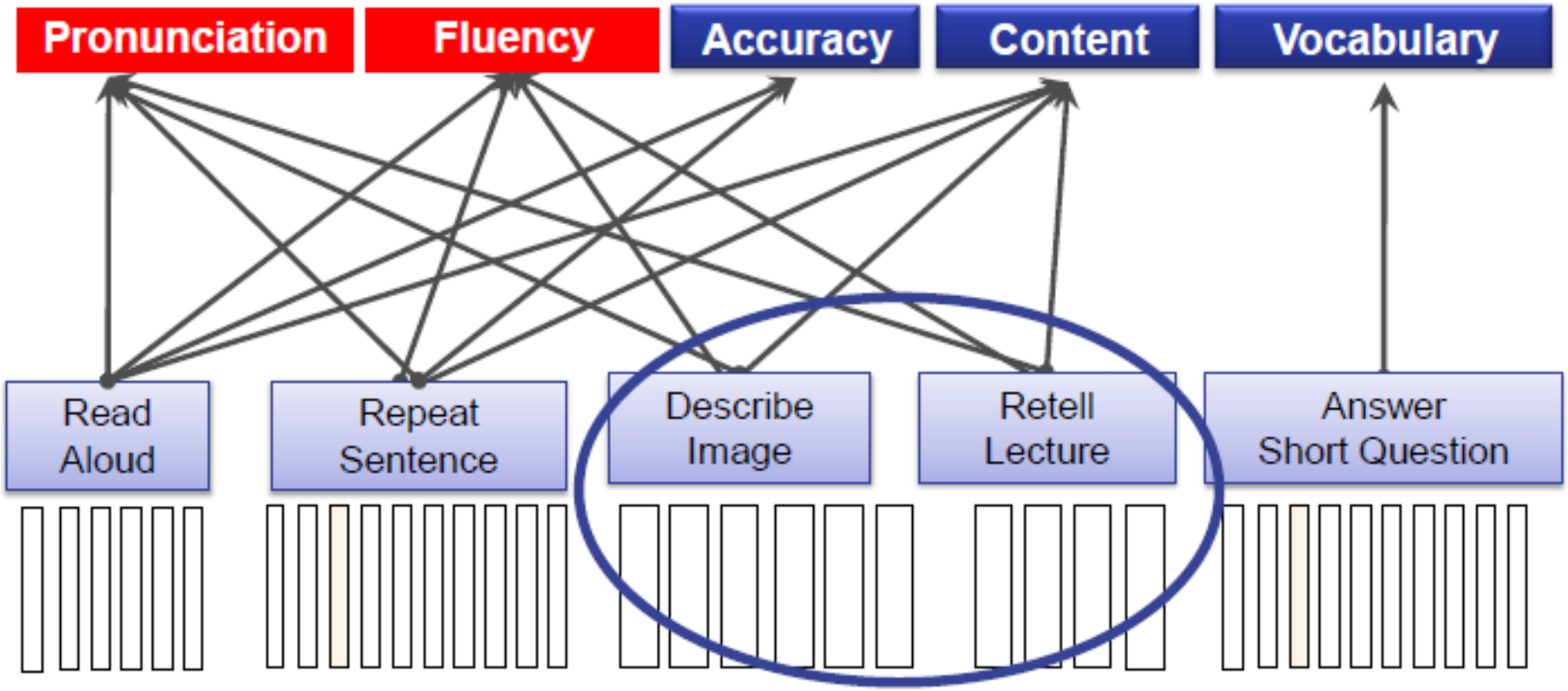
On behalf of Pearson

This presentation is based on research and implementation realized for the Pearson Test of English Academic (PTE Academic) and supported by Pearson's group managing PTE Academic

A variety item types, a high number of items and many score points addressing each skill

Skills Tested	Number of Item Types	Number of Items	Raw Score Points
Listening	11	40	123
Reading	9	27	101
Speaking	5	35	109
Writing	6	15	94

The example of Speaking



	Describe Image	Retell Lecture
Preparation time	25 secs	40 secs
Response time	40 secs	40 secs

The example of Speaking

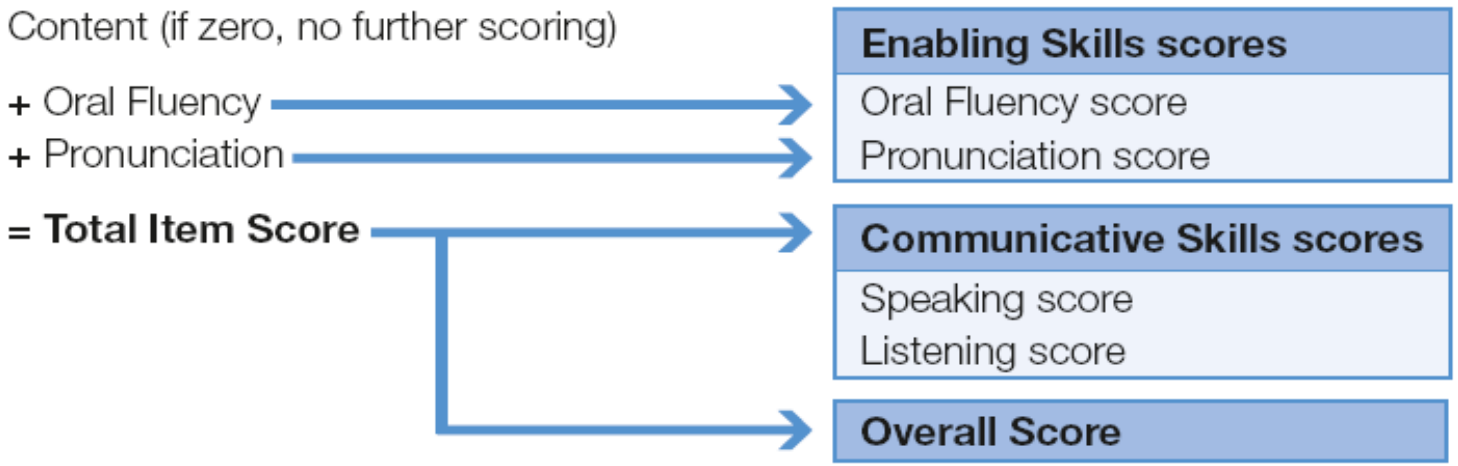
- **5 tasks:**
  - ~ 36 responses
  - ~ 8 minutes of speech
- **Input:**
  - Reading texts
  - Listening texts
  - Visual (non-linguistic)
- **Output:**
  - Prepared monologues
  - Short, real-time responses

Item scoring (partial credit)

Example: Integrated listening / speaking item

You will hear a lecture. After listening to the lecture, in 10 seconds, please speak into the microphone and retell what you have just heard from the lecture in your own words. You will have 40 seconds to give your response.

Your response is scored on:



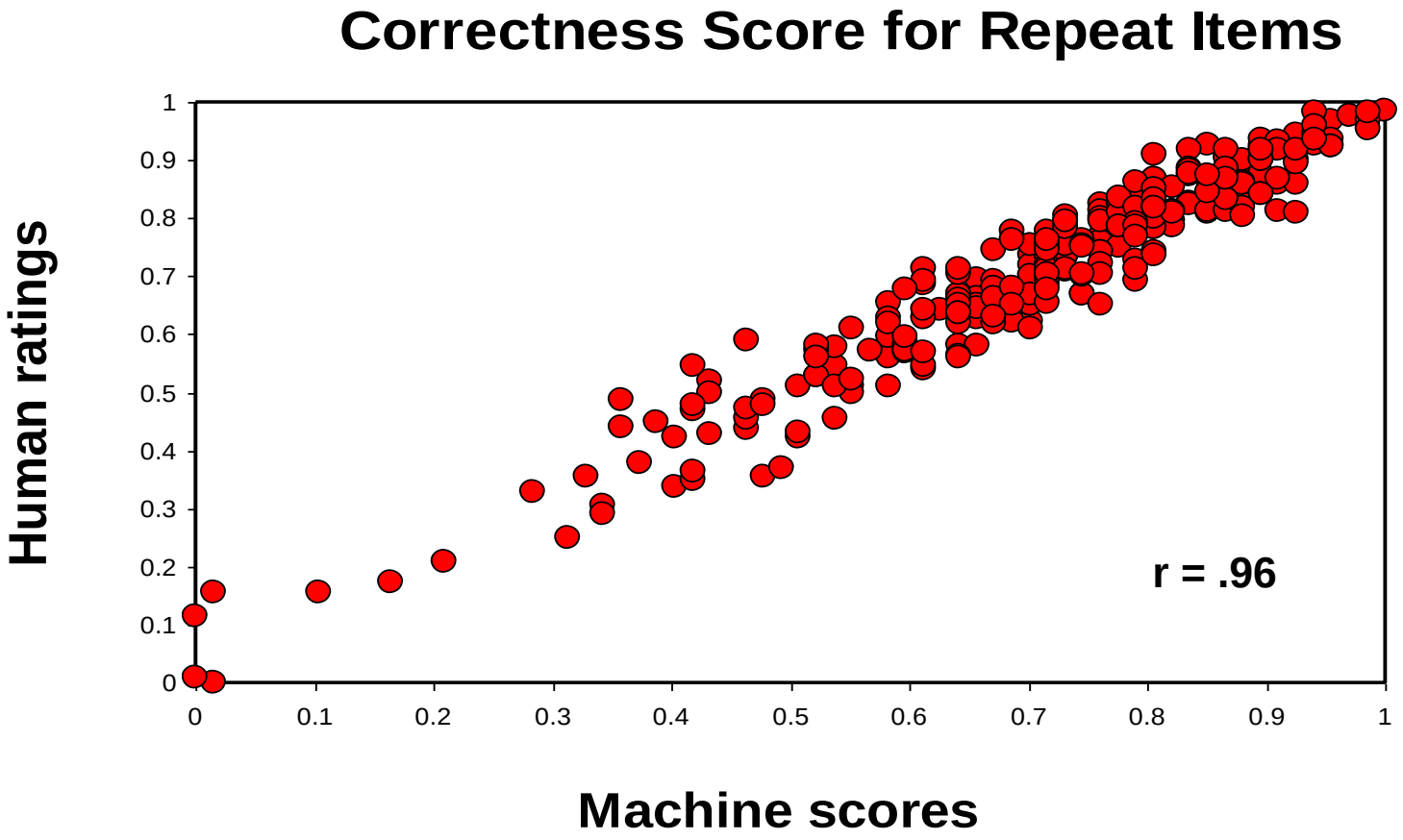
Latent Semantic Analysis (LSA)

- The content of what is written rather than just matching keywords
  - Texts are compared to each other in semantic space as similarity of responses is used to derive measures of quality, e.g. to match different phrases conveying the same meaning
  - Off-topic or highly unusual essays as well as plagiarism are detected
- The style, e.g. appropriate word choice, word and sentence flow, fluency, coherence, etc.
  - Detects larding of big words, non-standard language constructions, swear words, too long, too short …
- The mechanics, e.g. grammar, word usage, punctuation, spelling

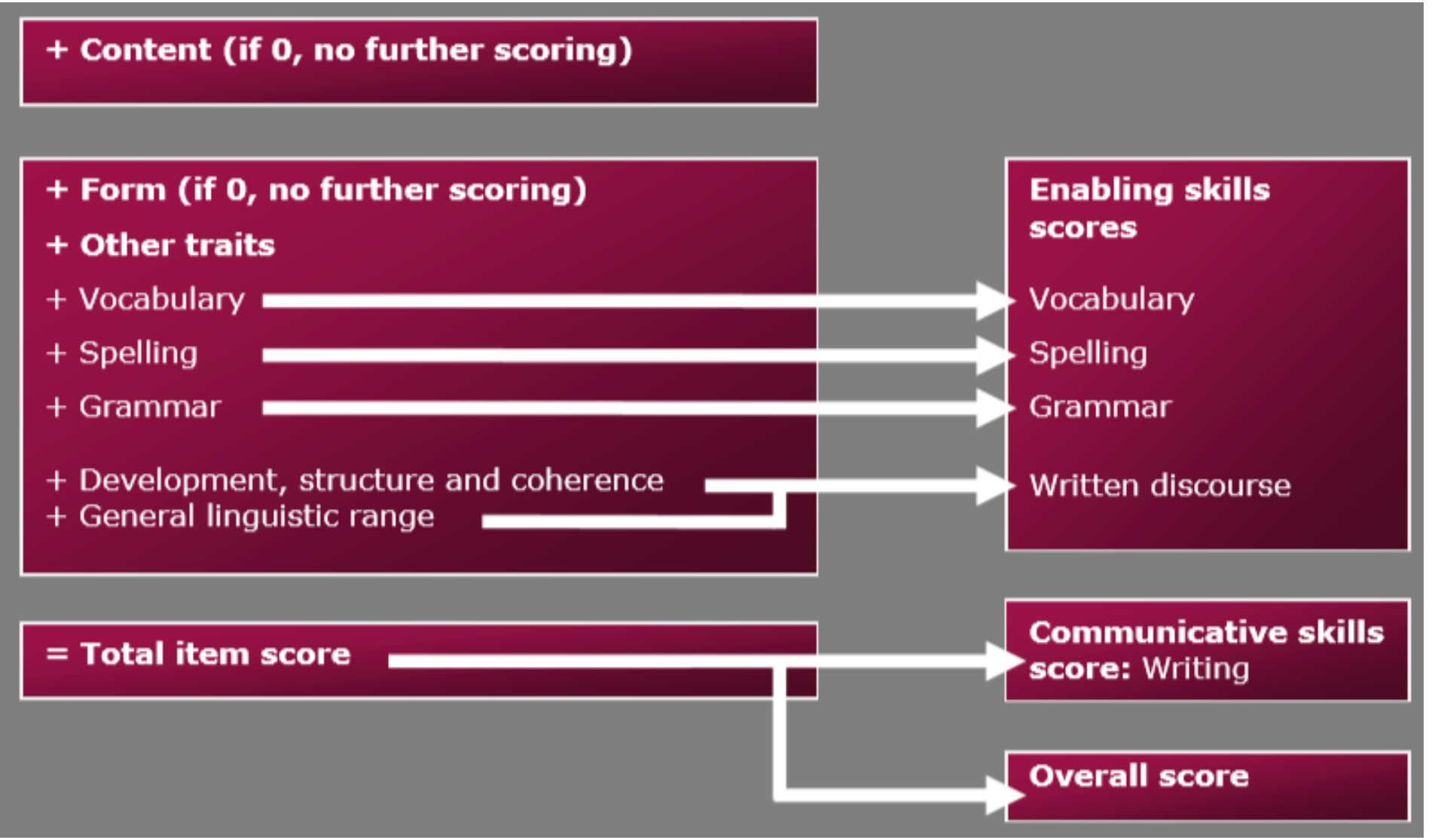
Validation of automated scoring

- **Two field tests** were done to validate the items and collect data for training the automated scoring engines (both written and spoken)
- Human scorers from US, Great Britain, Australia
- **Total of 2.6 million human ratings gathered**
  - ❑ 2 million for Speaking
  - ❑ 0.6 million for writing
- Several studies report **high agreement rates between automated scoring and human rating** (REF. ; Landauer, Laham, & Foltz, 2003; Landauer, Laham, Rehder, & Schreiner, 1997).

Machine to human correlations



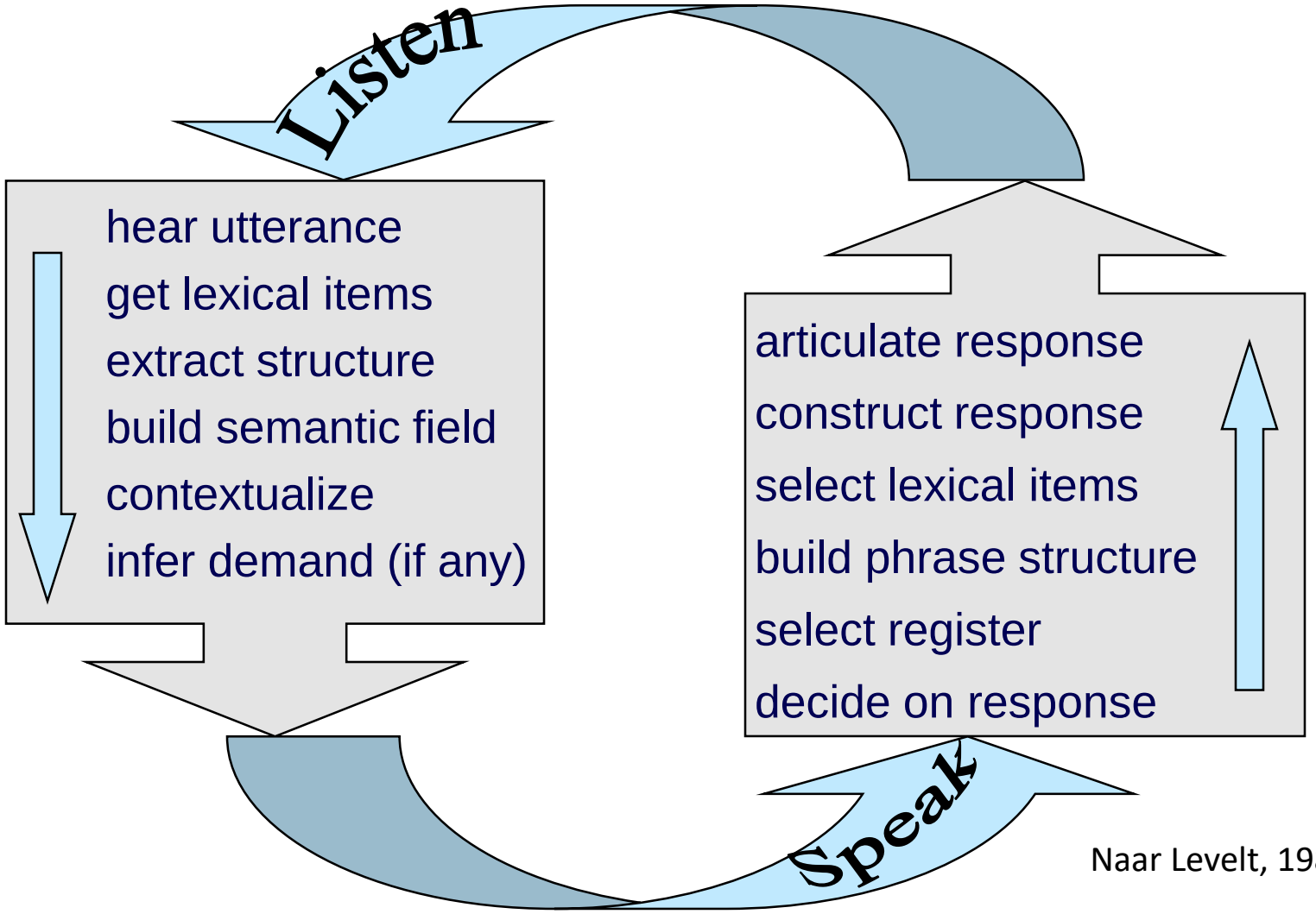
Scoring Essay



Automated Scoring: Training the Machine

To understand the way that the Pearson technology is “taught” to score spoken language, think about a person being trained by an expert rater to score speech samples during interviews. First, the expert rater gives the trainee rater a list of things to listen for in the test taker’s speech during the interview. Then the trainee observes the expert testing numerous test takers, and, after each interview, the expert shares with the trainee the score he or she gave the test taker and the characteristics of the performance that led to that score. Over several dozen interviews, the trainee’s scores begin to look very similar to the expert rater’s scores. Ultimately, one could predict the score the trainee would give a particular test taker based on the score that the expert gave.

Wat is being tested?



Naar Levelt, 1989

Intelligent Essay Assessor (IEA)

- IEA is trained individually for each prompt on 200-500 human scored responses
- IEA learns to score like the human markers by measuring different aspects of the responses
- IEA compares each new essay against all prescored essays to determine score

How Intelligent Essay Assessor (IEA) Works

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

IEA measures Content

- Semantic analysis measures of similarity to prescored responses, ideas, examples, ….

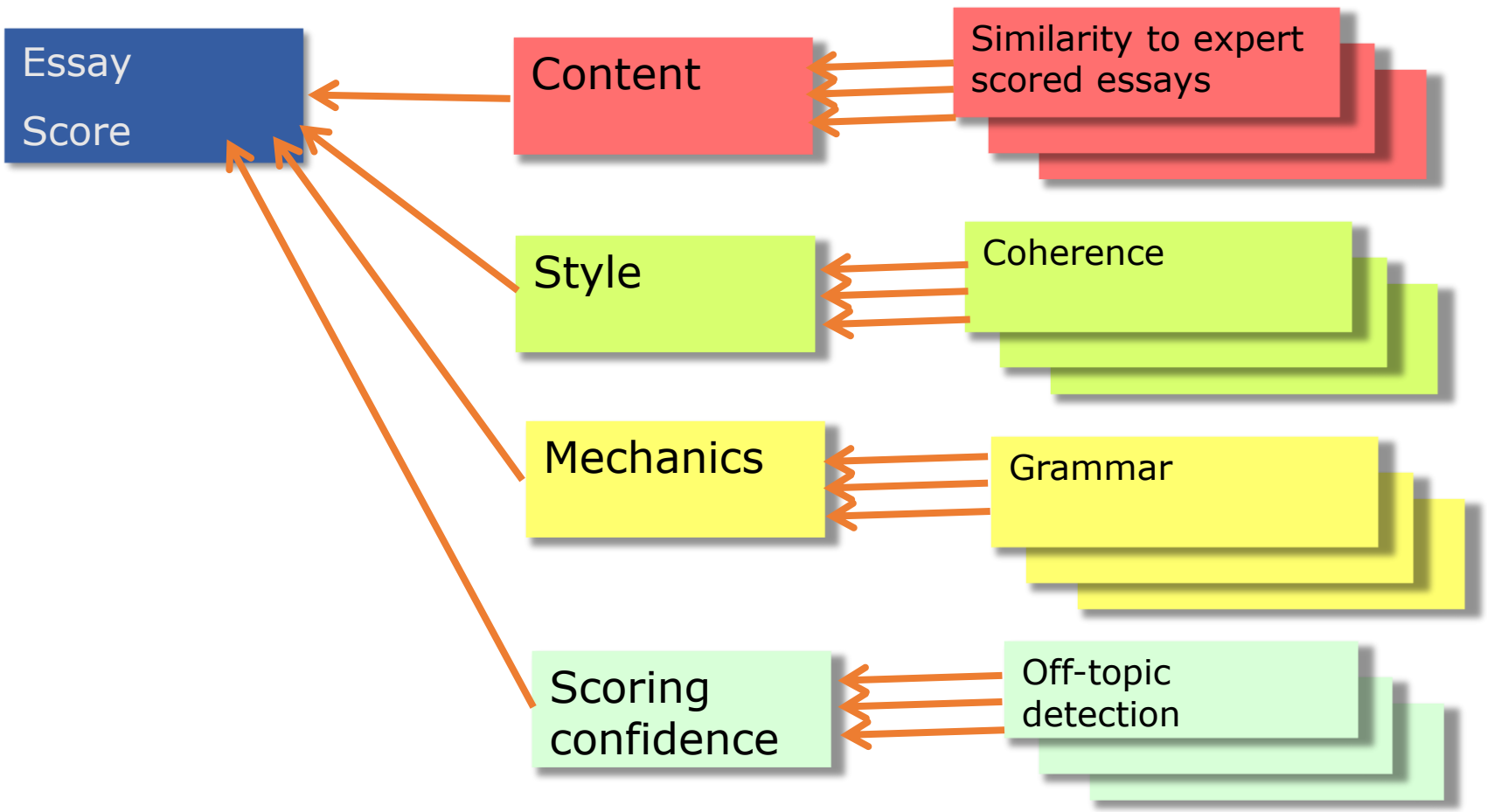
IEA measures Style

- Appropriate word choice, word and sentence flow, fluency, coherence, ….

IEA measures Mechanics

- Grammar, word usage, punctuation, spelling, …

Essay Scoring Process



Content-based scoring

Latent Semantic Analysis (LSA) to score CONTENT

a machine-learning technique using

- Linear algebra
- Enormous computing power

to capture the **meaning** of written English

See:

- *Surgery is often performed by a team of doctors.*
- *On many occasions, several physicians are involved in an operation.*

- LSA goes below surface structure to detect the latent meaning.
- LSA enables scoring the content of **what** is written rather than just matching keywords.
- The same technology is also widely used for search engines, spam detection, tutoring systems.

Latent Semantic Analysis background

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*"

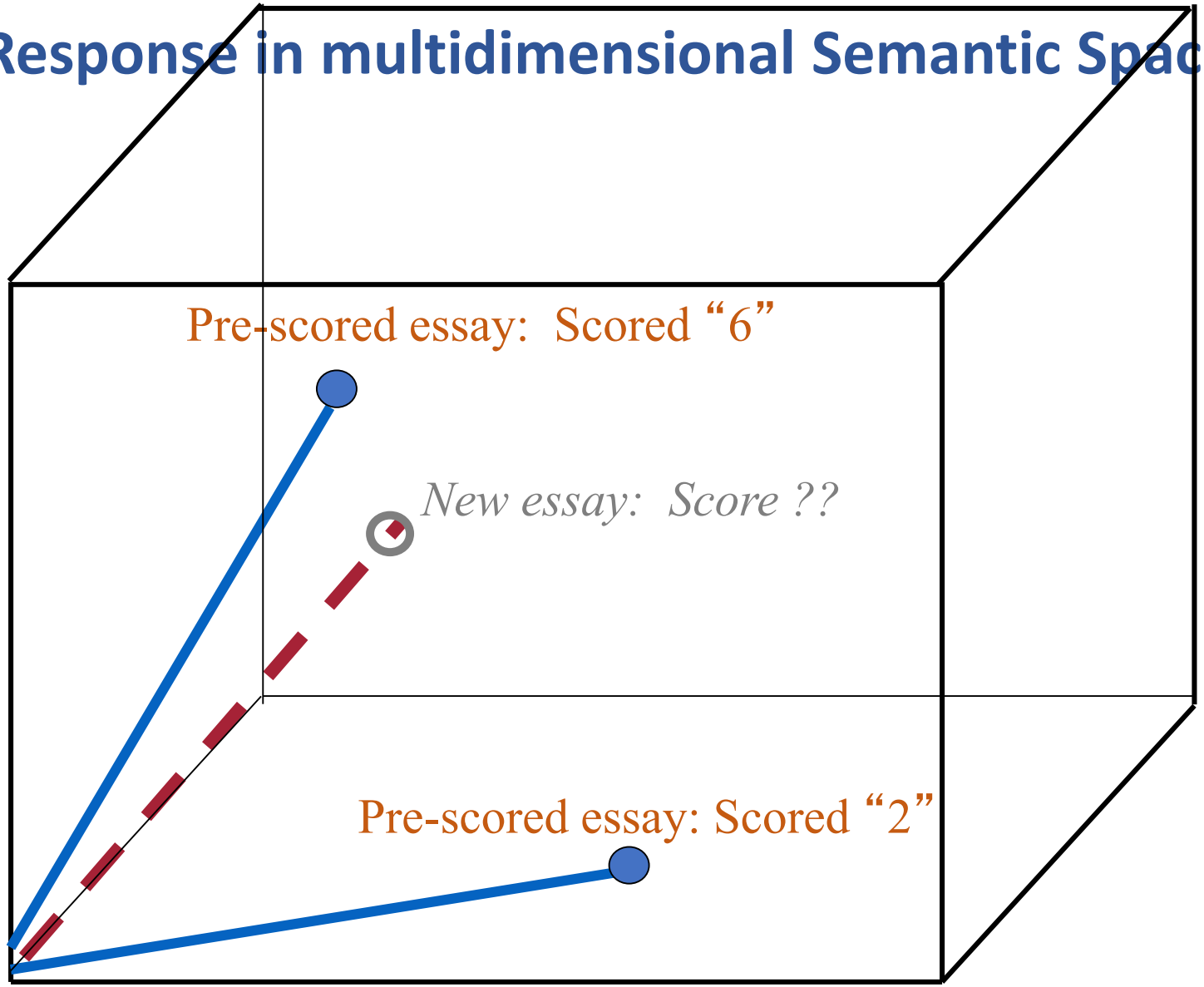
- Every word represented as a vector

Every paragraph represented as a vector

- M(Paragraph) = M(w1) + M(w2) + …

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

Placing a Response in multidimensional Semantic Space



KT Scoring Approach

Can score holistically, for content, and for individual writing traits

Content

Development

Response to the prompt

Effective Sentences

Focus & Organization

Grammar, Usage, & Mechanics

Word Choice

Development & Details

Conventions

Focus

Coherence

Progression of ideas

Style

Point of view

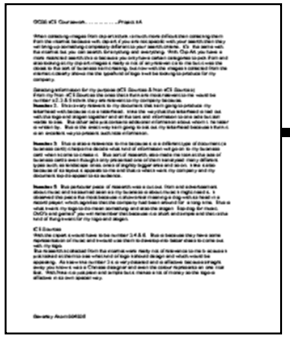
Critical thinking

Appropriate examples, reasons and other evidence to support a position

Sentence Structure

Skilled use of language and accurate and apt vocabulary

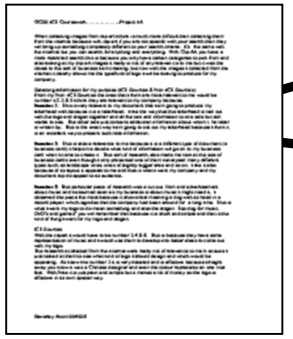
Development



Human Scorers

System is “trained” to predict human scores

Validation



Expert human ratings

Very highly correlated

Machine scores

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”
- Can be done in other languages
- Plagiarism

Reliability and Validity

IEA has been tested on millions of essays

- 4th grade through college, medical school, professional school, standardized tests, job applications, military

Generally agrees with a single human reader as often as 2 human readers agree with each other

The more skilled the human readers, the better the agreement

Consistent, Objective, Immediate

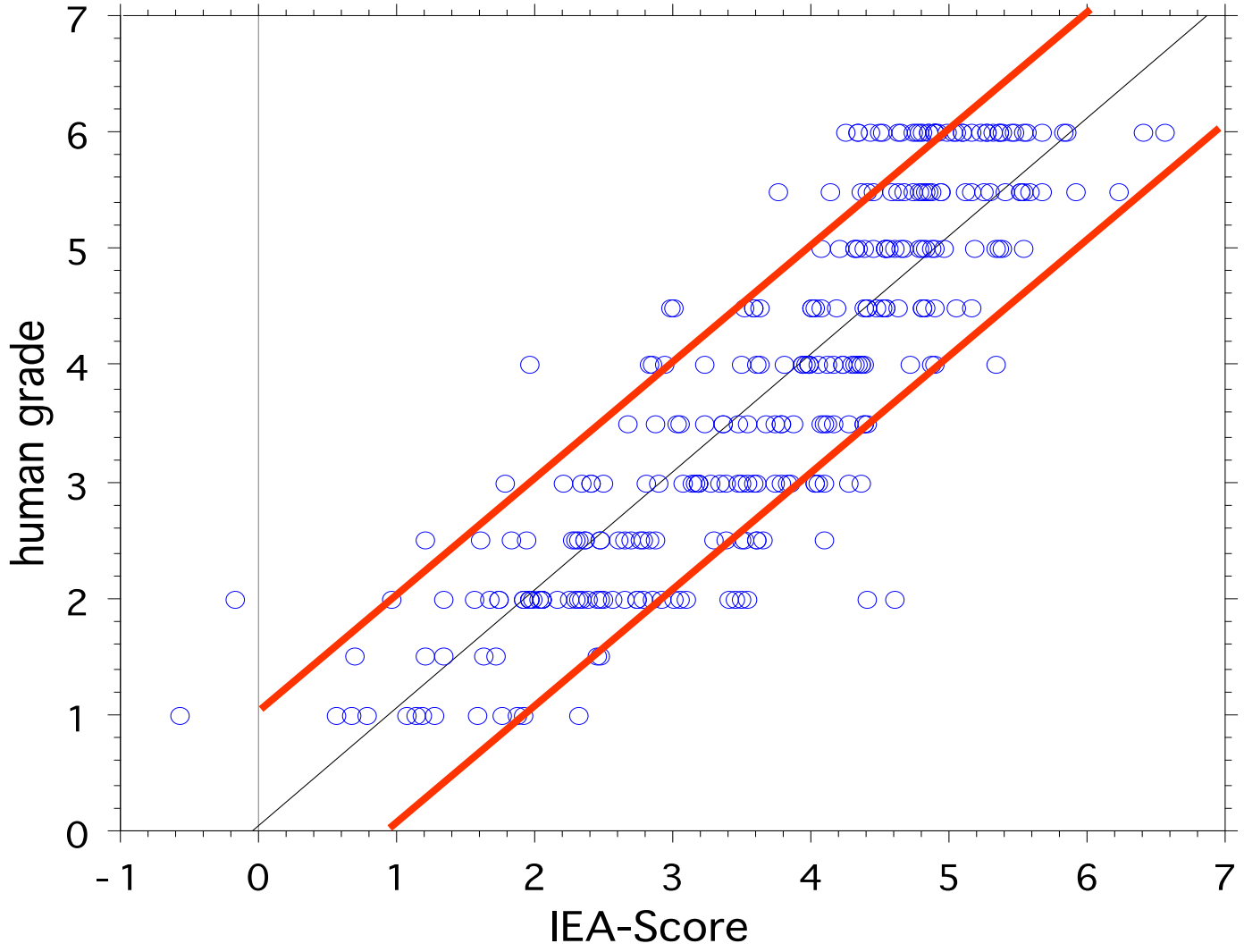
Catches off-topic and other irregular essays

Reliability of Essay Scoring

- 99 diverse prompts; 4th-12th grade students
- Scoring developed using essays with scores by operational readers of a major testing company.
- Trained on essays, tested on others

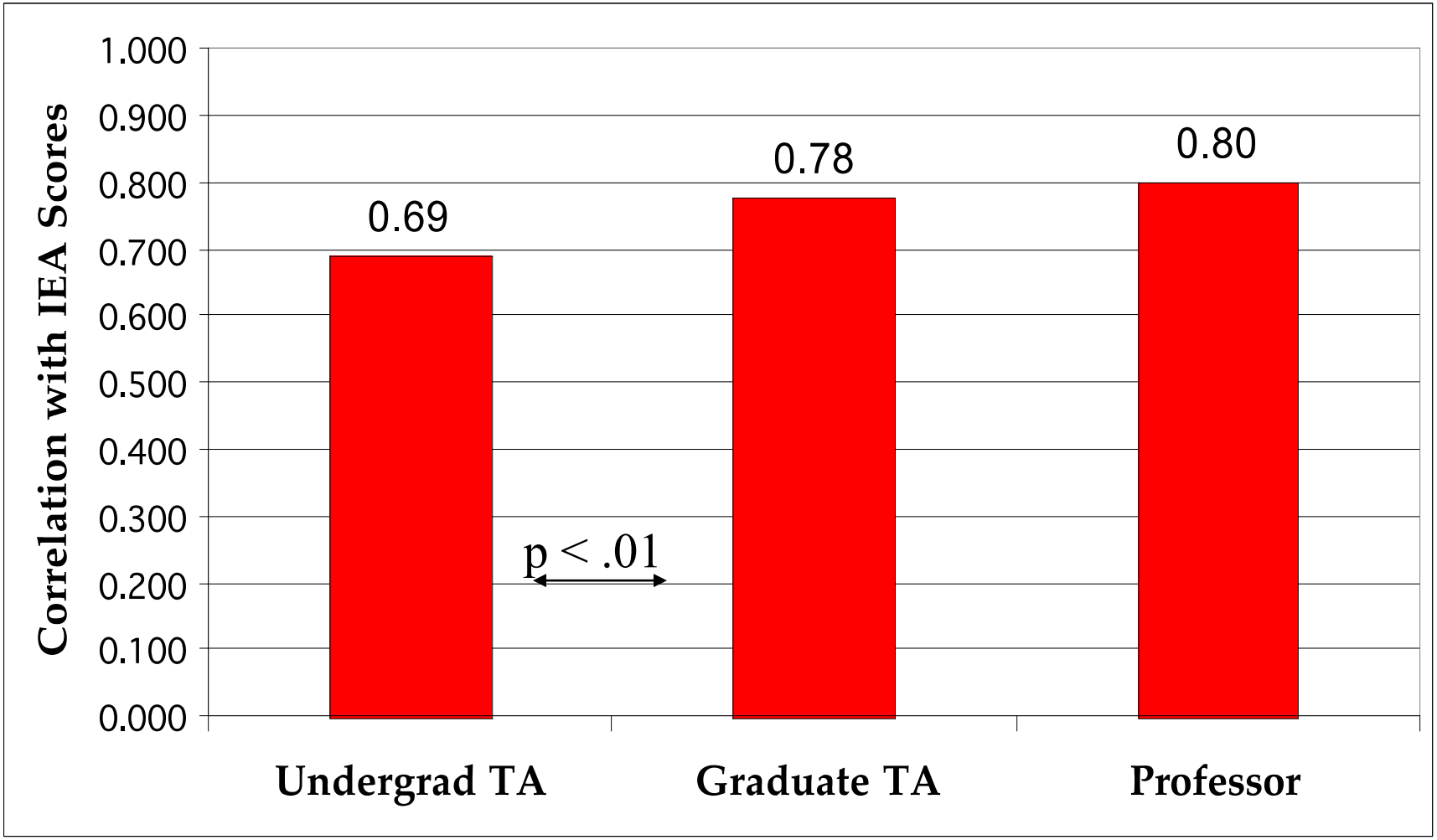
Measure	Automated Scoring to human raters (min, mean, max)	Human raters to human raters (min, mean, max)
Correlation	.76 **.88** .95	.74 **.86** .95
Exact score agreement	50% **63%** 81%	43% **63%** 87%
Exact + adjacent agreement	91% **98%** 100%	87% **98%** 100%

Scattergram for GMAT 1 Test Set



External validity of IEA

IEA agrees with better trained scorers



Creative Essays

Prompt: “Usually the basement door was locked but today it was left open…”

900 Narrative Essays

Scored by an international testing organization

IEA agrees with human readers as well as the human readers agree with each other (correlation of 0.9)

Validity of IEA predicting school grade of student

	human grader scores	Intelligent Essay Assessor scores
Correct school grade	66%	74%

How is that possible? Unbelievable!

Have a look here Wikipedia bots:

ClueBot NG, as the bot is known, resides on a computer from which it sallies forth into the vast encyclopedia to detect and clean up vandalism almost as soon as it occurs.

Interwiki bots

- link articles on the same subject in different languages
- flag potential copyright violations and other irregularities for human review

And what about this:

Thursday 10 Jul 2014 6:00 am

The Associated Press (AP) news agency has just started using technology that will automatically generate thousands of financial reports without the need of reporters.

In AP's case, Wordsmith will write thousands of earnings stories that would not have otherwise existed. Wordsmith operates at a speed and scale humans cannot match. Our technology frees journalists to do more interesting work.

And this

A "robot scientist", Adam is able to perform independent experiments to test hypotheses and interpret findings without human guidance, removing some of the drudgery of laboratory experimentation.

Adam is capable of:

- hypothesizing to explain observations
- devising experiments to test these hypotheses
- physically running the experiments using laboratory robotics
- interpreting the results from the experiments
- repeating the cycle as required

While researching yeast-based functional genomics, Adam became the first machine in history to have discovered new scientific knowledge independently of its human creators.