Automated Scoring of Written Open Responses

John H.A.L. de Jong	Language Testing
Peter W. Foltz	Knowledge Technologies
Ying Zheng	Language Testing

Click here

Talk Overview

How written item scoring works

How well it works

Some existing applications

Considerations, limitations, and future directions

Why Automated Scoring?

Accuracy

- As accurate as skilled human graders

Speed

- Get reports back more quickly

Consistency

- A score of 3 today is a score of 3 tomorrow

Objectivity

- Knows when it doesn‛t know

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, ….

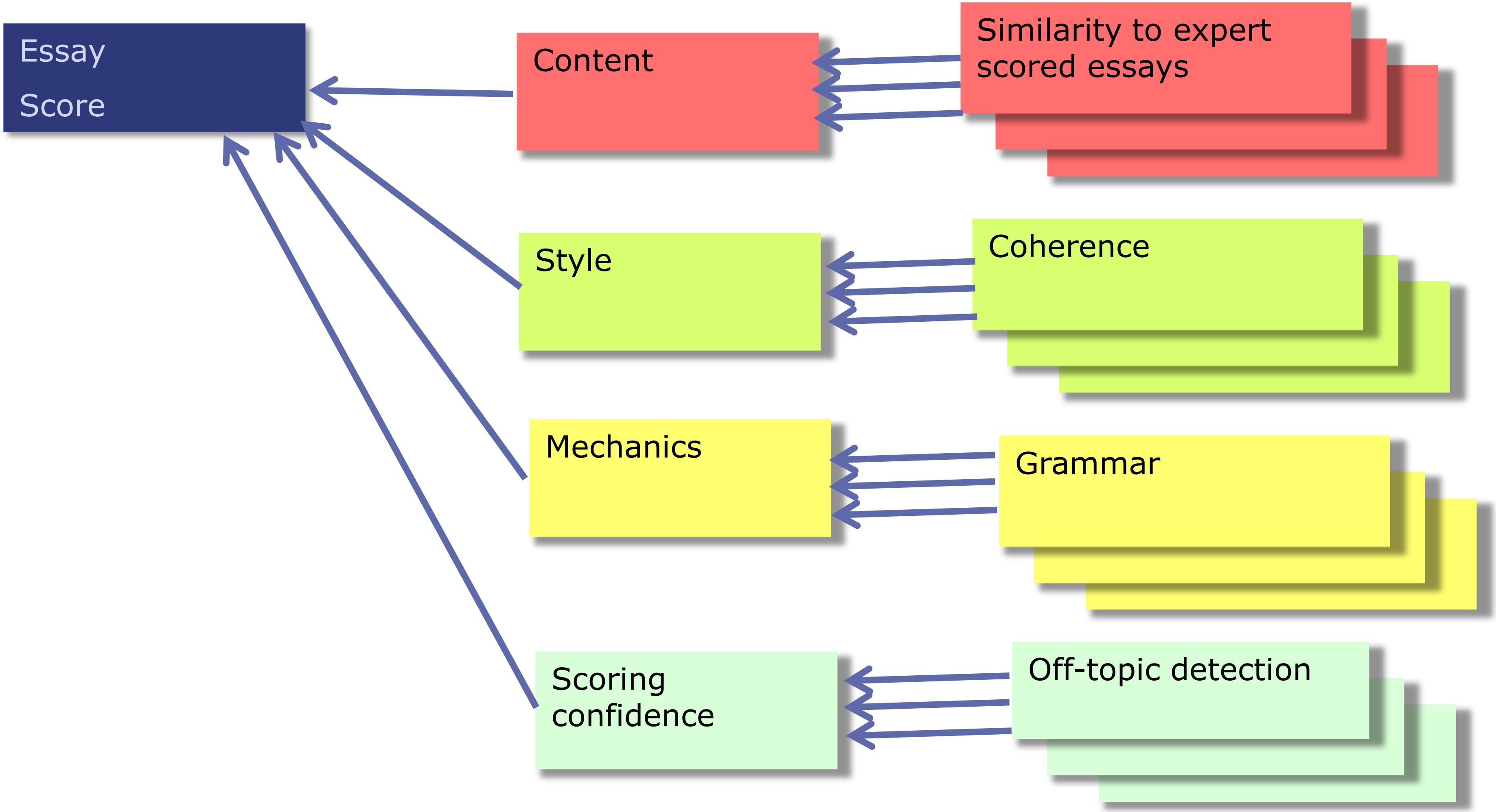
IEA measures Style

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

IEA measures Mechanics

- Grammar, word usage, punctuation, spelling, …

Essay Scoring Process



Essay Score

Content

Similarity to expert scored essays

Style

Coherence

Mechanics

Grammar

Scoring confidence

Off-topic detection

Content-based scoring

Content of essays is scored using Latent Semantic Analysis (LSA), a machine-learning technique using

- Linear algebra
- Enormous computing power

to capture the **meaning** of written English.

The following two sentences have not a single word in common:

- *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. The machine knows that those two sentences have almost the same meaning.

LSA enables scoring the content of what is written rather than just matching keywords.

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*"

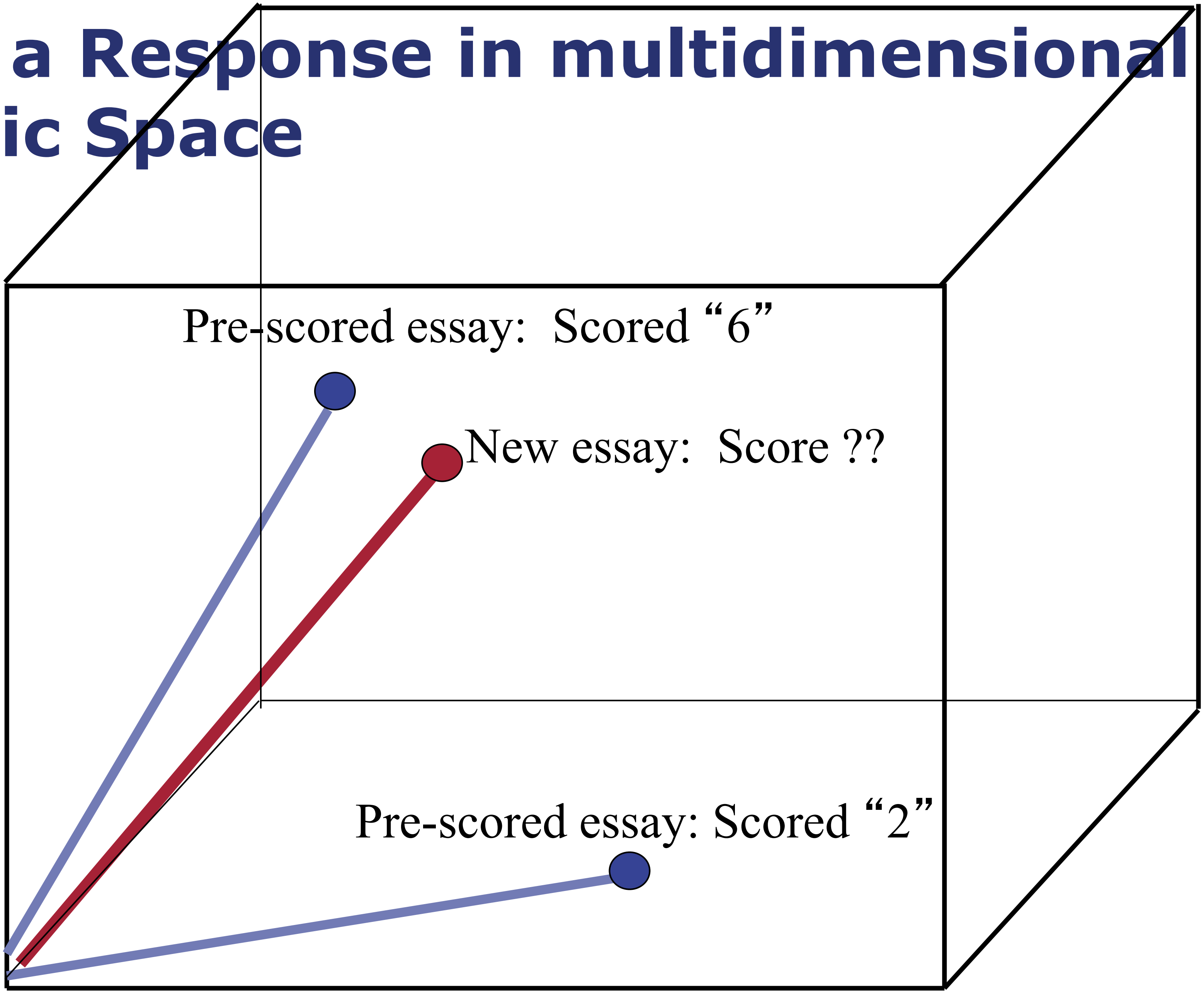
- Every word represented as a vector

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

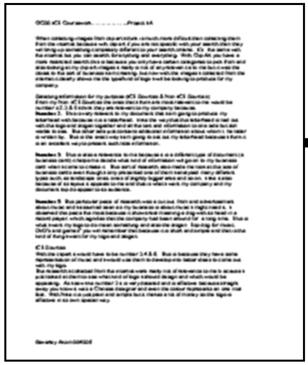
Critical thinking

Appropriate examples, reasons and other evidence to support a position.

Sentence Structure

Skilled use of language and accurate and apt vocabulary

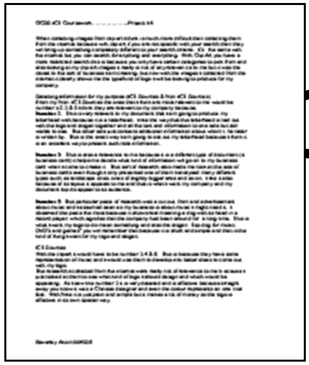
Development



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

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

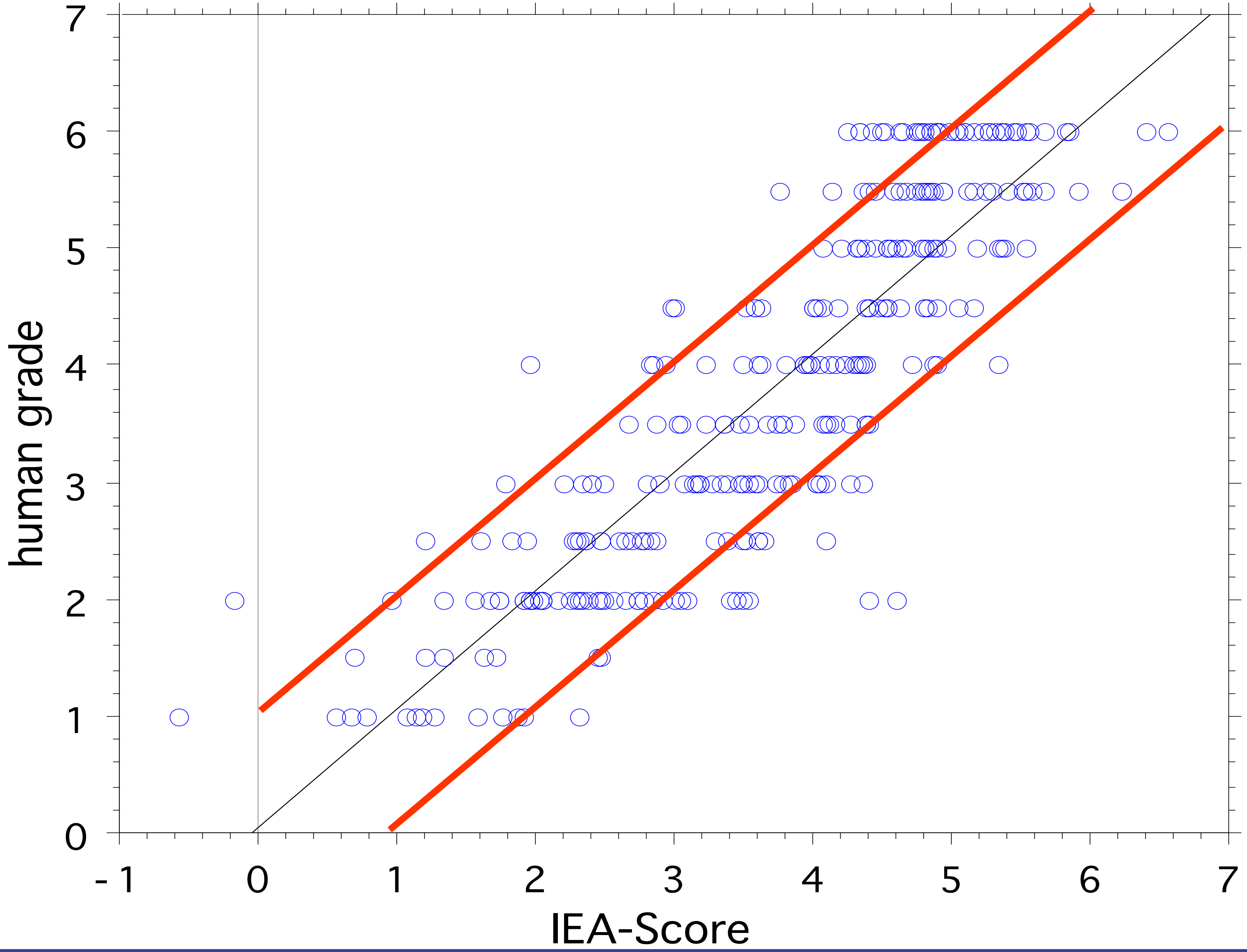
Catches off-topic and other irregular essays

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

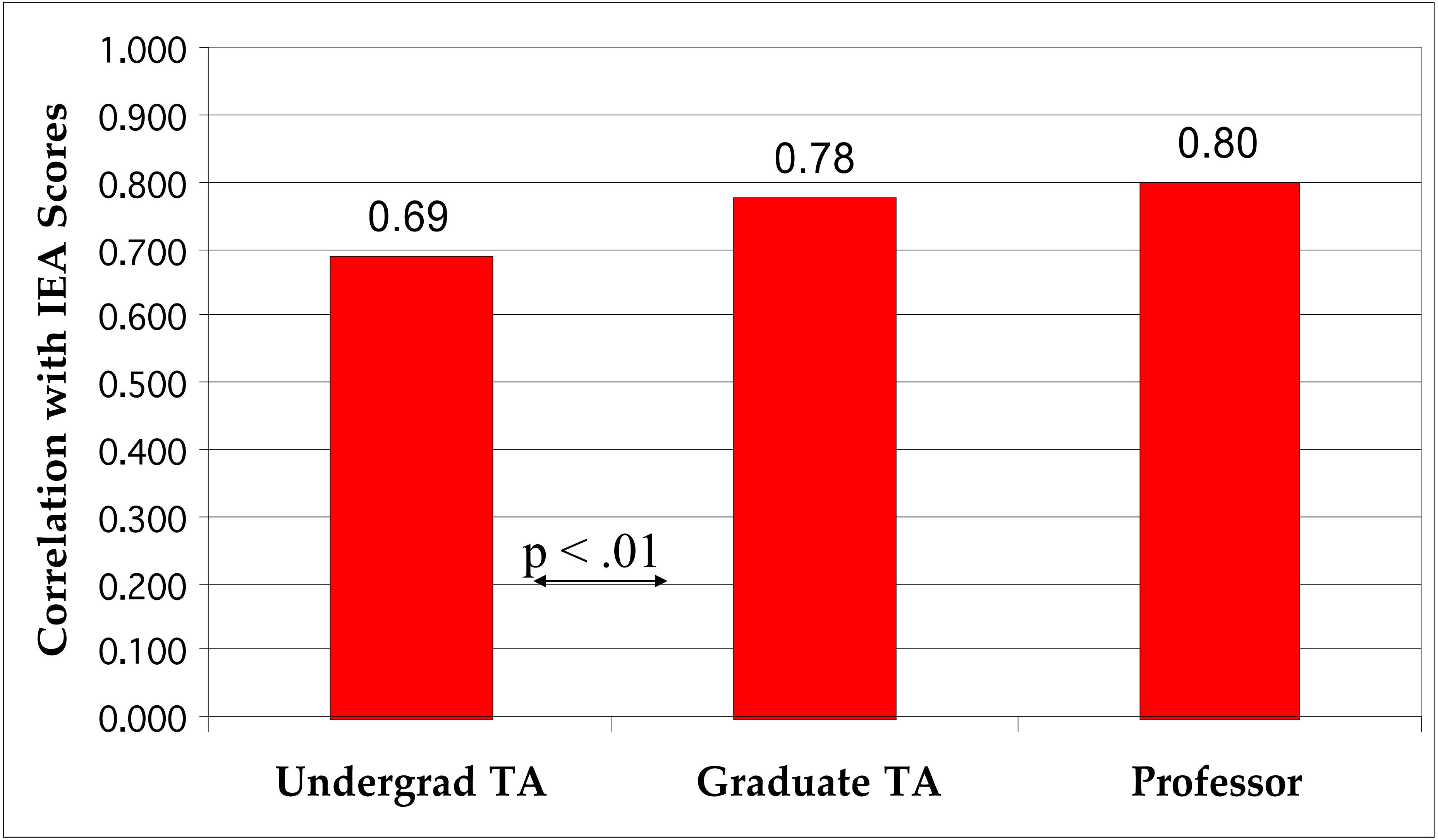
Measure	Automated Scoring to human raters (min, mean, max)	Human raters to human raters (min, mean, max)
Correlation	.76 **.88** .95	.74 **.86** .95
Exact score agreement	50% **63%** 81%	43% **63%** 87%
Exact + adjacent agreement	91% **98%** 100%	87% **98%** 100%

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%

IEA In Operation

State Assessments

- South Dakota

Writing Practice

- Prentice Hall; Holt McDougal; Kaplan
- WriteToLearn
- Writing Coach

Higher Ed Placement/Evaluation

- ACCUPLACER
- Council for Aid to Education (CAE)
- Pearson Test of English - Academic

Brief Constructed Responses

5 to 25 word responses

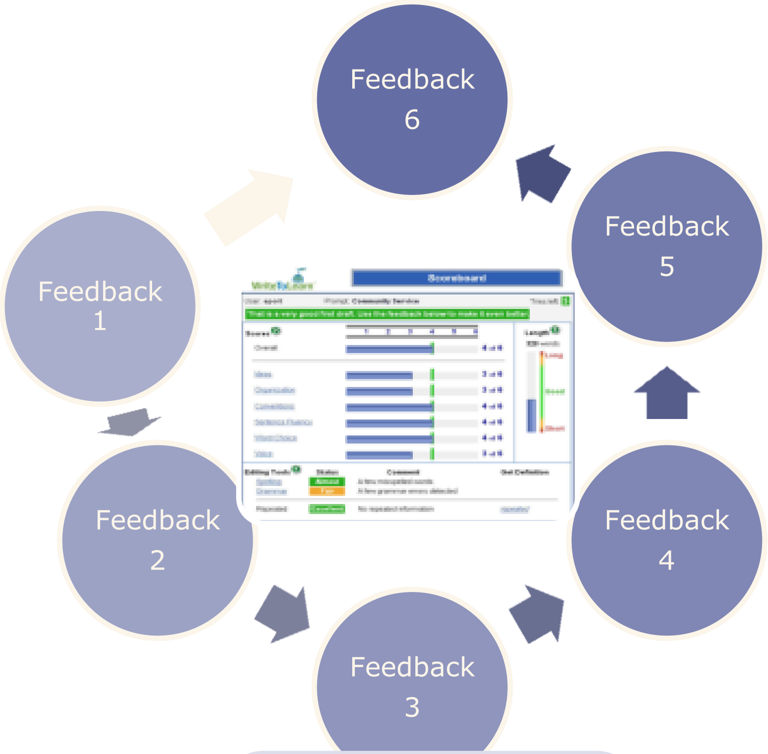
Used for scoring content knowledge, comprehension more than expression

Can be more difficult to score than “long” responses

- Training Data: 500 responses across the score points
- Automatically identify and correct misspellings
- Use a combination of IEA/LSA and a statistical classifier to analyze the responses
- Learn to distinguish among the score categories based on the examples
- Test Data: 500 additional responses used to evaluate performance

Incorporating automated writing assessment in the classroom

Students write an essay or a summary to a prompt or a text assigned by the teacher



Prompt: Community Service

That is a very good first draft. Use the feedback below to make it even better.

Scores	
Overall	4 of 6
Ideas	3 of 6
Organization	3 of 6
Conventions	4 of 6
Sentence Fluency	4 of 6
Word Choice	4 of 6
Voice	3 of 6

Editing Tools	Status	Comment
Spelling	Almost	A few misspelled words
Grammar	Fair	A few grammar errors detected
Repeated	Excellent	No repeated information

Students get immediate and accurate feedbacks while they write

Teachers check writing scores of classroom and individuals and monitor the progress

WriteToLearn

Online tool for building writing skills and developing reading comprehension

- Writing instruction through practice
- Reading comprehension through summarization
- Immediate, automated evaluation with targeted feedback
    - Six traits of writing
    - Summary quality and missing information
    - Grammar, spelling, redundancy, off-topic sentences, …

Studies of WriteToLearn components compared to control groups

- Significantly better comprehension and writing from two weeks of use (Wade-Stein & Kintsch, 2004)
- Increased content scores compared to controls (Franzke et al., 2005)
- Improved gist skills on standardized comprehension test; Scores as reliably as human raters

Writing Coach

Interactive writing coach

Feedback on paragraph

- Topic development, focus, organization,

Feedback on overall essay

- sentence variety, word choice, six traits of writing

Paragraph Level Feedback

Topic Focus: A rating of how well the sentences of the paragraph support the topic.

Topic Development: A rating of how well the topic is developed over the course of the paragraph. Does the paragraph have too many ideas, too few, or just the right amount?

Sentence Variety Length: Do the sentences of the paragraph vary appropriately in length?

Sentence Beginnings Variety: Do the beginnings of each sentence vary sufficiently?

Sentence Structure Variety: Do the structures of the sentences vary appropriately?

Transitions: Select transition words can be identified

Vague Adjectives can be identified

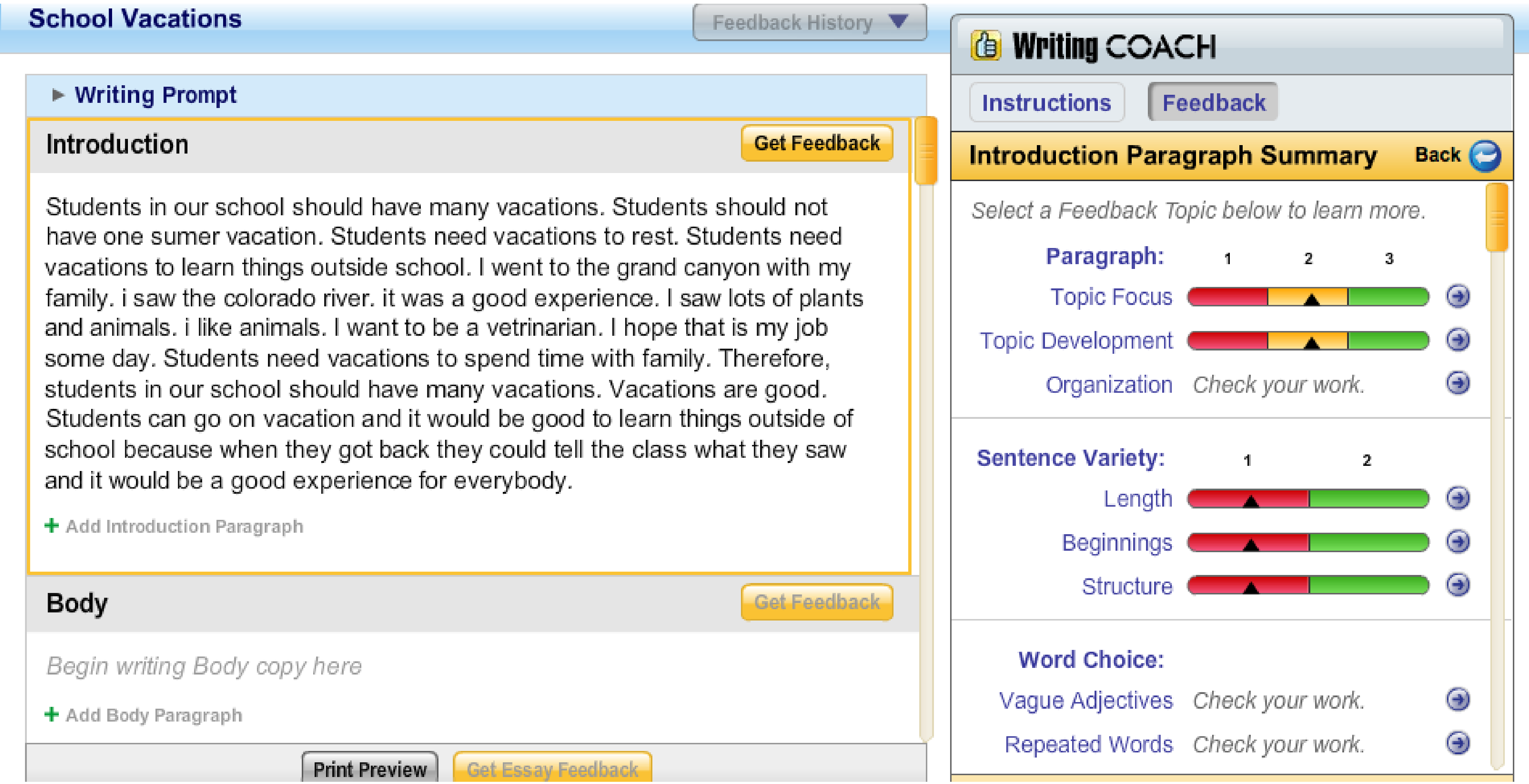
Repeated Words can be identified

Pronouns can be identified

Spelling errors can be identified and suggested corrections made

Grammar errors can be identified and suggested corrections made

Redundant sentences can be identified



Students in our school should have many vacations. Students should not have one sumer vacation. Students need vacations to rest. Students need vacations to learn things outside school. I went to the grand canyon with my family. i saw the colorado river. it was a good experience. I saw lots of plants and animals. i like animals. I want to be a vetrinarian. I hope that is my job some day. Students need vacations to spend time with family. Therefore, students in our school should have many vacations. Vacations are good. Students can go on vacation and it would be good to learn things outside of school because when they got back they could tell the class what they saw and it would be a good experience for everybody.

Begin writing Body copy here

Introduction Paragraph Summary

Select a Feedback Topic below to learn more.

Paragraph: 1 2 3

Topic Focus

Topic Development

Organization *Check your work.*

Sentence Variety: 1 2

Length

Beginnings

Structure

Word Choice:

Vague Adjectives *Check your work.*

Repeated Words *Check your work.*

General comments about automated scoring

Scoring is based on the collective wisdom of many skilled human scorers

- How humans have scored similar responses, even if different words, different approaches

Is as accurate or more accurate than humans

Perfectly consistent, purely objective, and completely impartial

Fast: <1 – 3 seconds per response

New Research Directions for Scoring Enhancements

Improved validation methods

- More diverse ways of detecting “unscorables”

More fine-grained analysis within essays

- For feedback and overall performance
- Argument detection and evaluation
- Response to texts

Questions?