Pearson

Practical Considerations for Implementing Automated Scoring of Writing for Large-Scale Use

Peter W. Foltz
Pearson and University of Colorado

AI for

Rapid g

Genera

This isn

OCTOBER 25 2017 - 1:00PM

‘Robots can’t read essays’

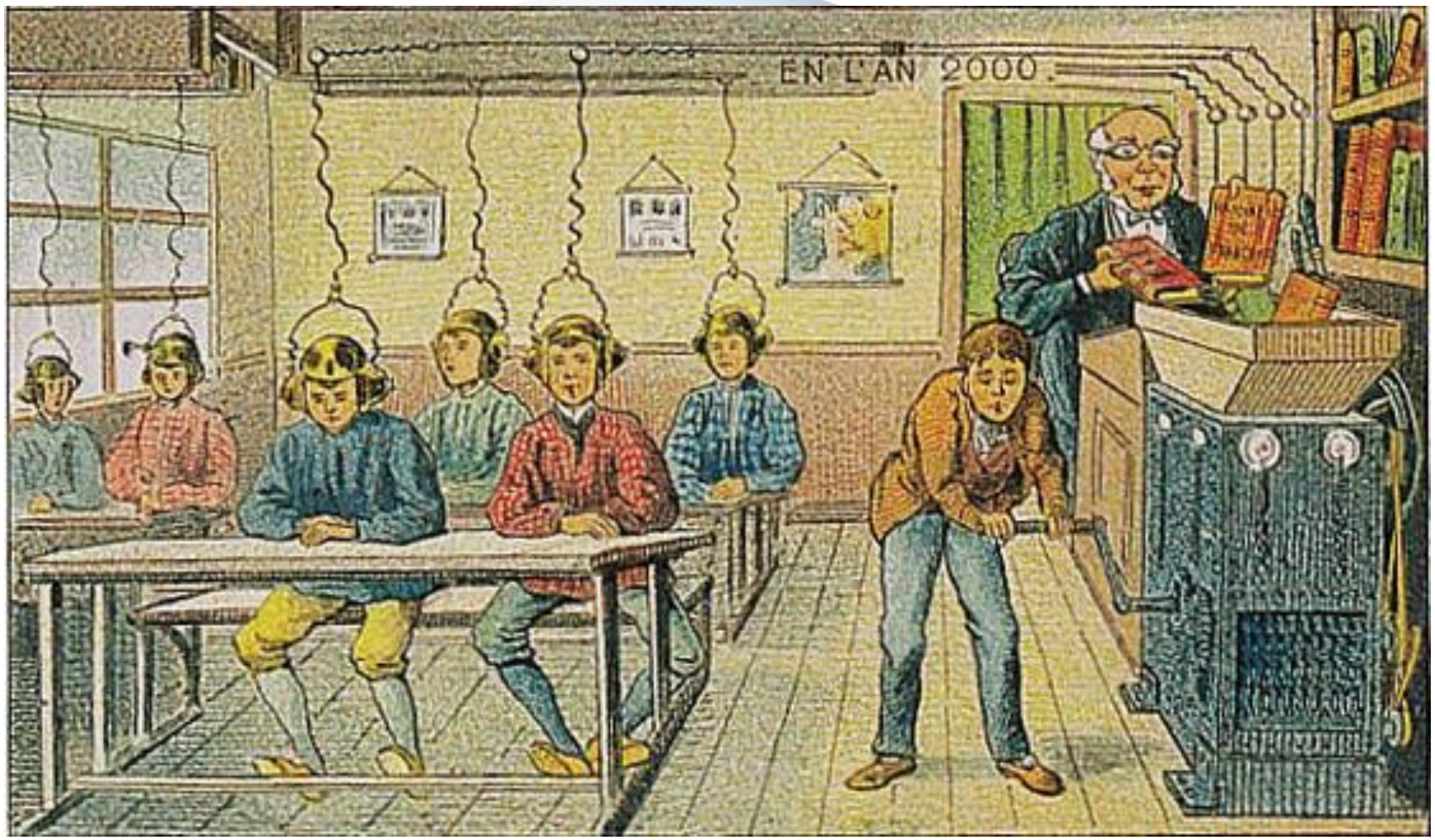
Daily Liberal

Local News

SHARE TWEET

humans”

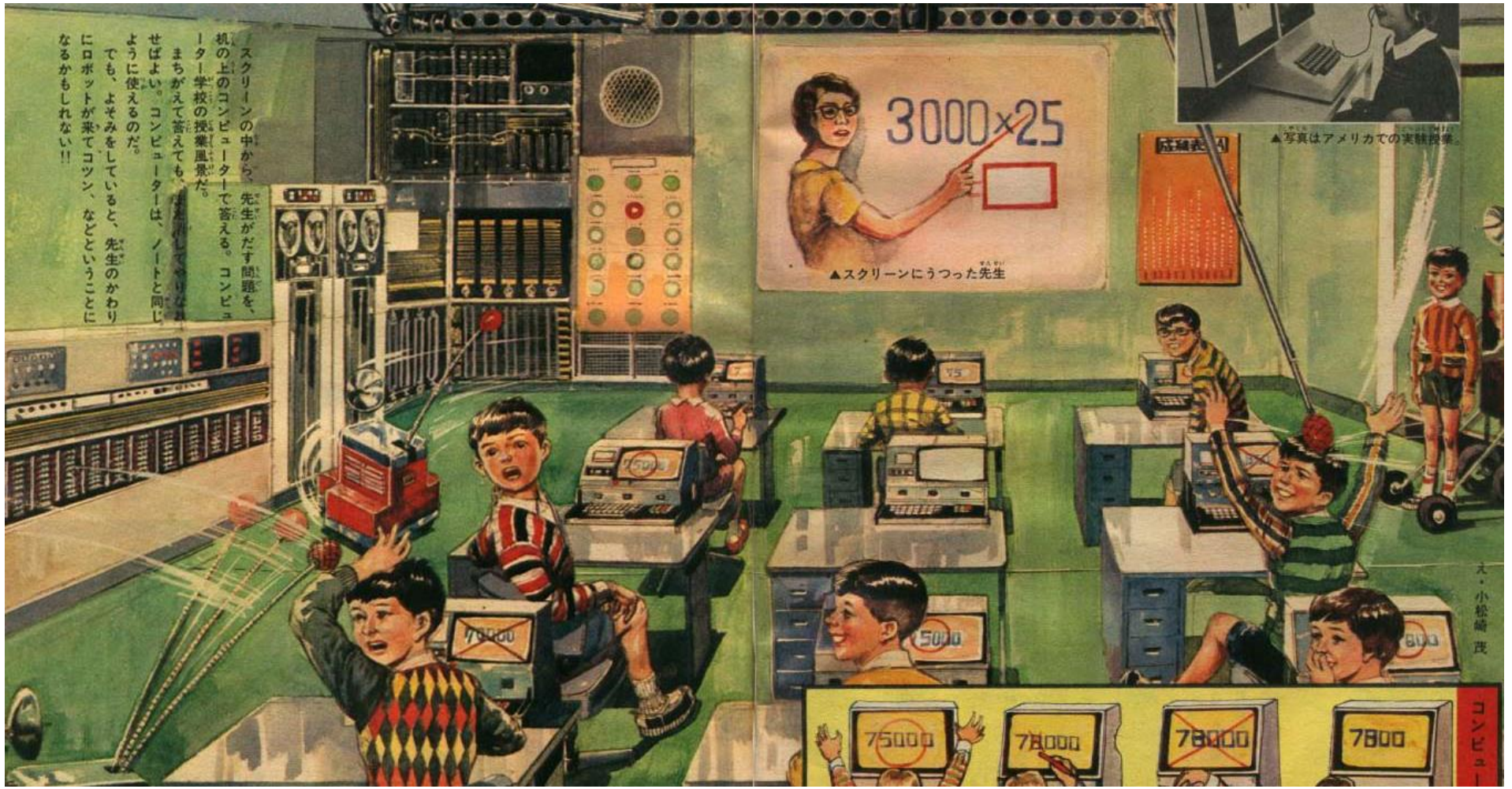
–



France “In the year 2000” (1910)

PUSH-BUTTON EDUCATION Tomorrow's schools will be more crowded; teachers will be correspondingly fewer. Plans for a push-button school have already been proposed by Dr. Simon Ramo, science faculty member at California Institute of Technology. Teaching would be by means of sound movies and mechanical tabulating machines. Pupils would record attendance and answer questions by pushing buttons. Special machines would be "geared" for each individual student so he could advance as rapidly as his abilities warranted. Progress records, also kept by machine, would be periodically reviewed by skilled teachers, and personal help would be available when necessary.

The future of education (1958)



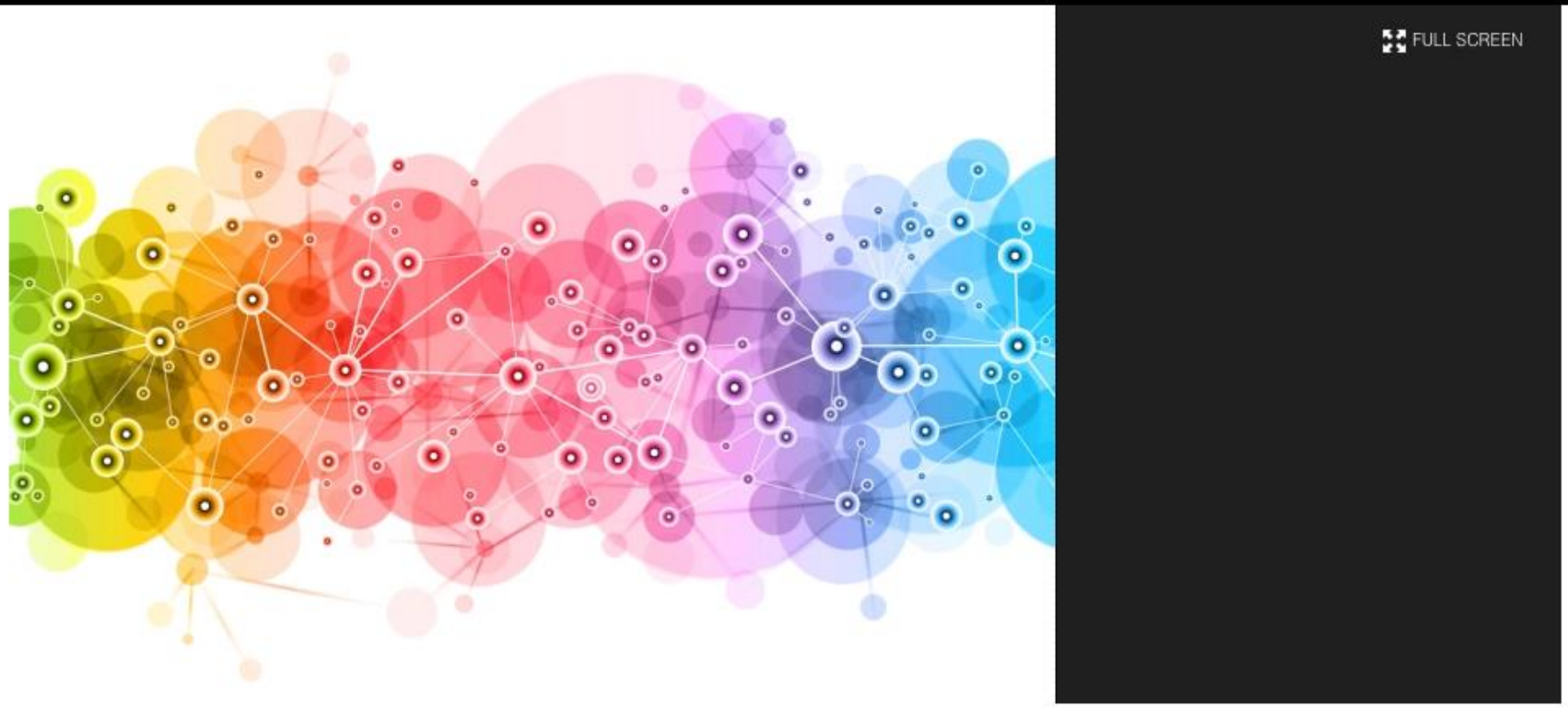
Future of education Japan (1969)


THINK
$24,000
Who is Stoker?
((FOR ONE WELCOME OUR NEW COMPUTER OVERLORDS))
$ 1,000
$77,147
Who is Bram Stoker?
$ 17,973
$21,600
WHO IS BRAM STOKER?
$5600

Driverless cars could reduce traffic fatalities by up to 90%, says report

They're set to have one of the biggest impacts on public health ever.

BEC CREW 1 OCT 2015



Automating big-data analysis

System that replaces human intuition with algorithms outperforms 615 of 906 human teams.

Larry Hardesty | MIT News Office
October 16, 2015

Press Inquiries

RELATED

AI in Education

- Not many in Education and Assessment understand AI
- Not many in AI understand Education or Assessment

How do we foster innovative uses that improve student learning and teacher effectiveness?

How do we educate practitioners/users/consumers about use of AI in education?

How do we ensure best practices across the diverse fields?

Best practices for AI in Assessment

Understand the assumptions that go into the modeling

Keep humans in the loop

Automated scoring of writing

Immediacy & Efficiency

- Evaluate responses in seconds
- Give students and teachers instant feedback
- Can be integrated in large-scale summative assessment
- Scores for writing traits content, ideas, word choice, organization, …
- Grammar, Spelling, …

Accuracy

Consistency, Objectivity

Can detect off-topic, inappropriate and “odd” responses

A stored record of a student’s effort

Pearson

Understand the assumptions underlying automated scoring

Building an automated scoring model

Creating a scoring model:

Inferring human scorer behavior

- Computer learns background knowledge of the domain by “reading” a large amount of text (corpus)
- Computer is trained on a large sample of human-scored essays
- Computer uses machine learning learns to associate combination of language features with scores (or types of feedback) for particular writing traits

Similar to Evidence Centered Design (e.g. Mislevy, Behrens, Dicerbo, & Levy, 2012)

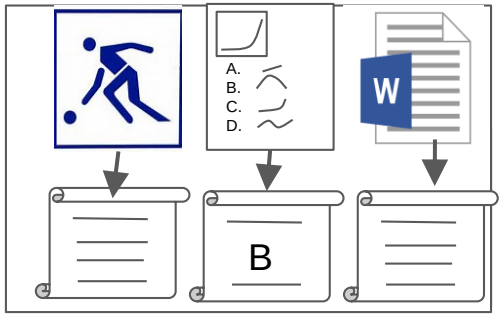
- Evidence model defines the writing constructs of interest
- Construct representation in the scoring rubric
- But … provide explicit linkage between the features being scored and the construct.
- Support validity argument for how combined scoring features represent construct of interest

In a physical world, the teacher has lots of these constructs in private mental models.

Example: *Is a misspelling part of the construct for writing organization? Should students get feedback for it? Does it affect their scores?*

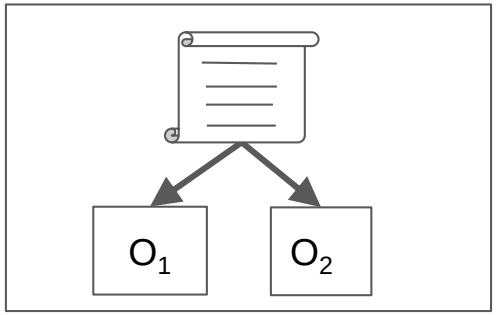
In digital systems we need to be more explicit and precise.

Task Model

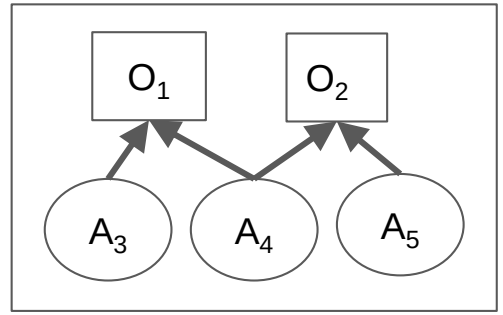


Scoring

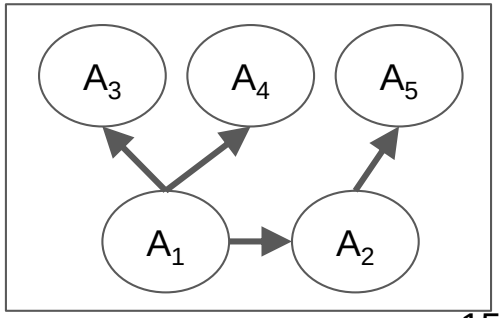
Evidence Model

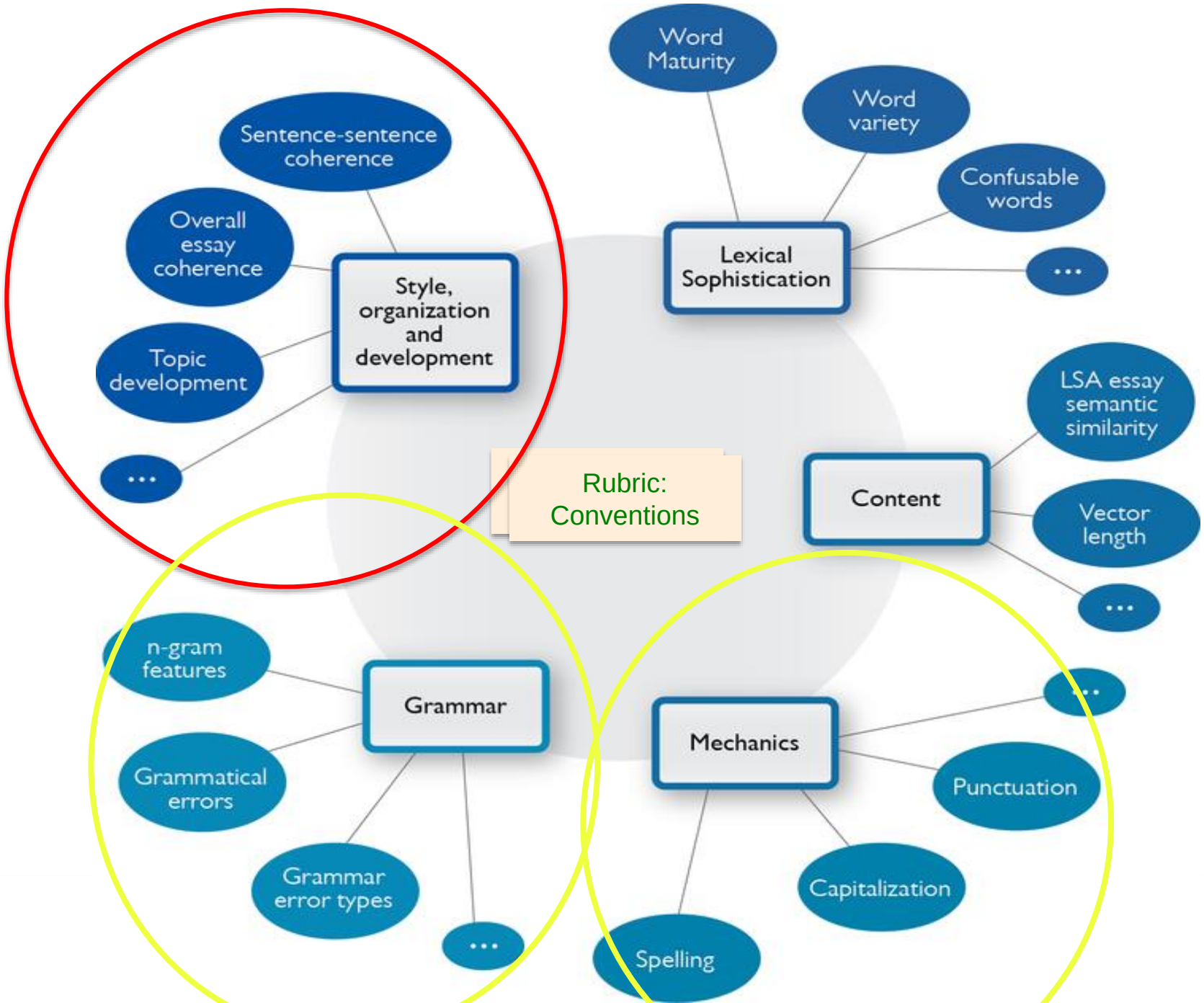


Weighting



Student Model

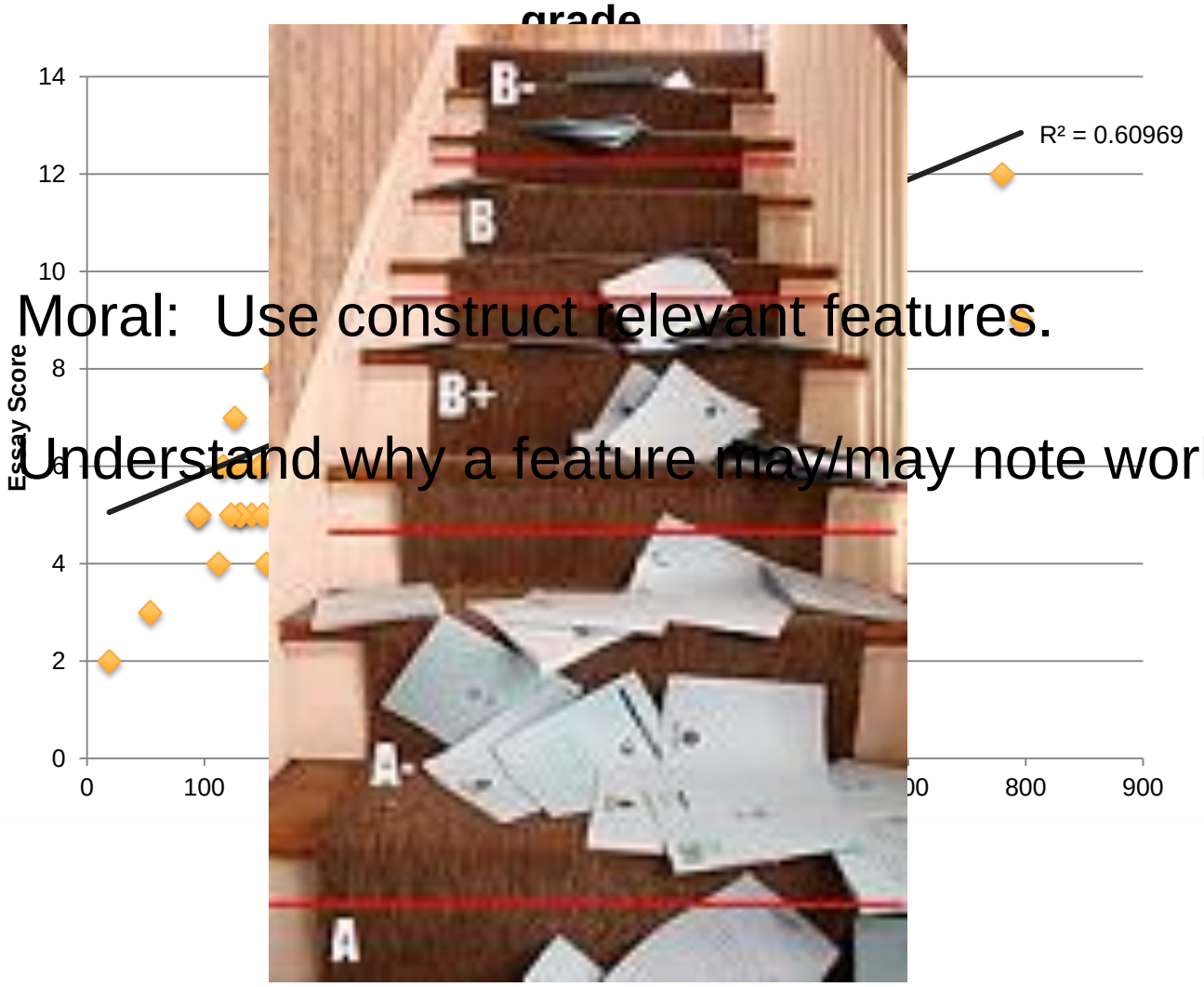



Word Maturity
Word variety
Confusable words
...
Lexical Sophistication
Sentence-sentence coherence
Overall essay coherence
Topic development
...
Style, organization and development
LSA essay semantic similarity
Vector length
...
Content
Rubric: Conventions
n-gram features
Grammatical errors
Grammar error types
...
Grammar
...
Punctuation
Capitalization
Spelling
Mechanics

Pearson

Building automated scoring: A thought experiment:

Why does throwing essays down the stairs work?



Moral: Use construct relevant features.

Understand why a feature may/may note work

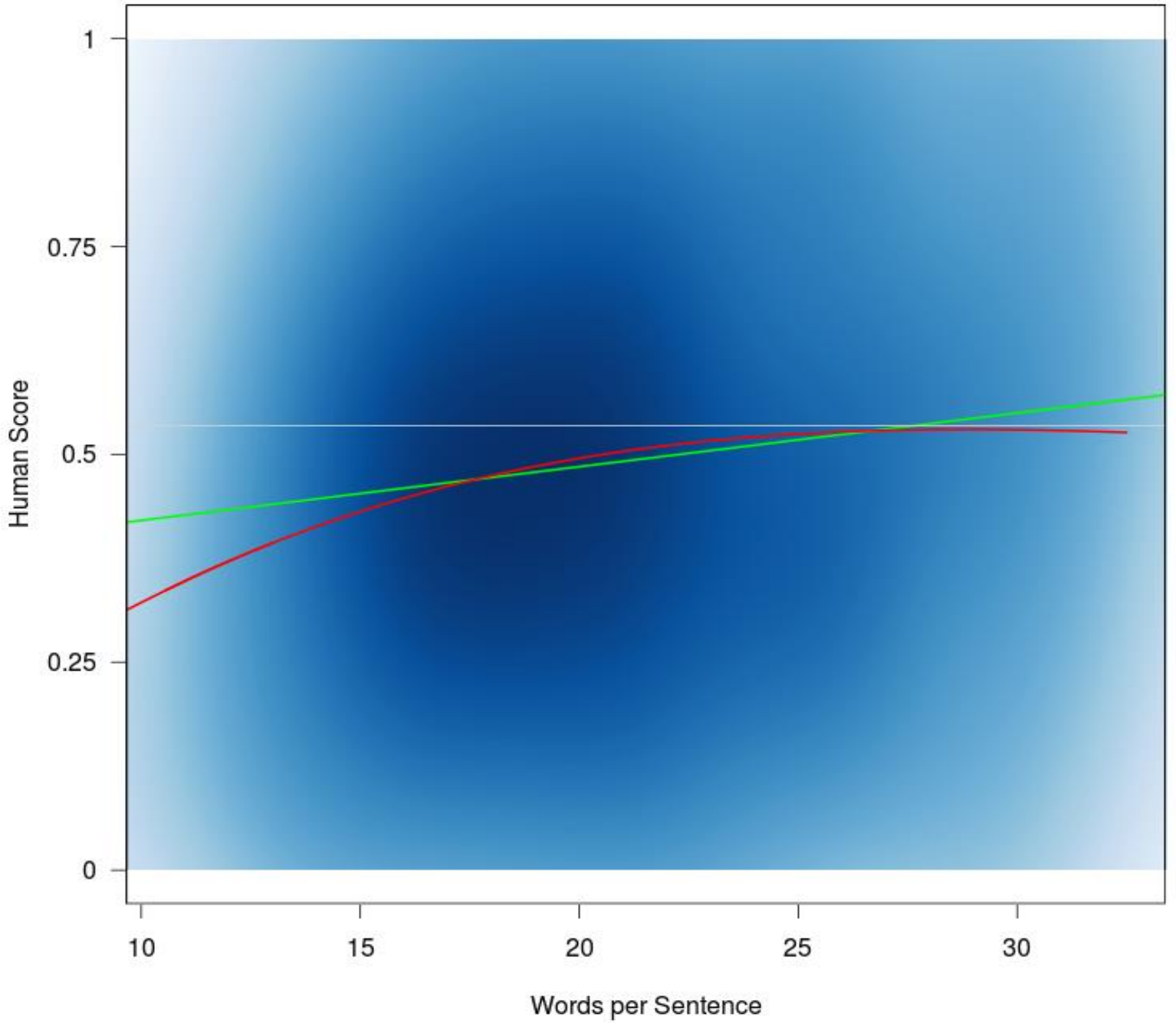
Non-linearities in features

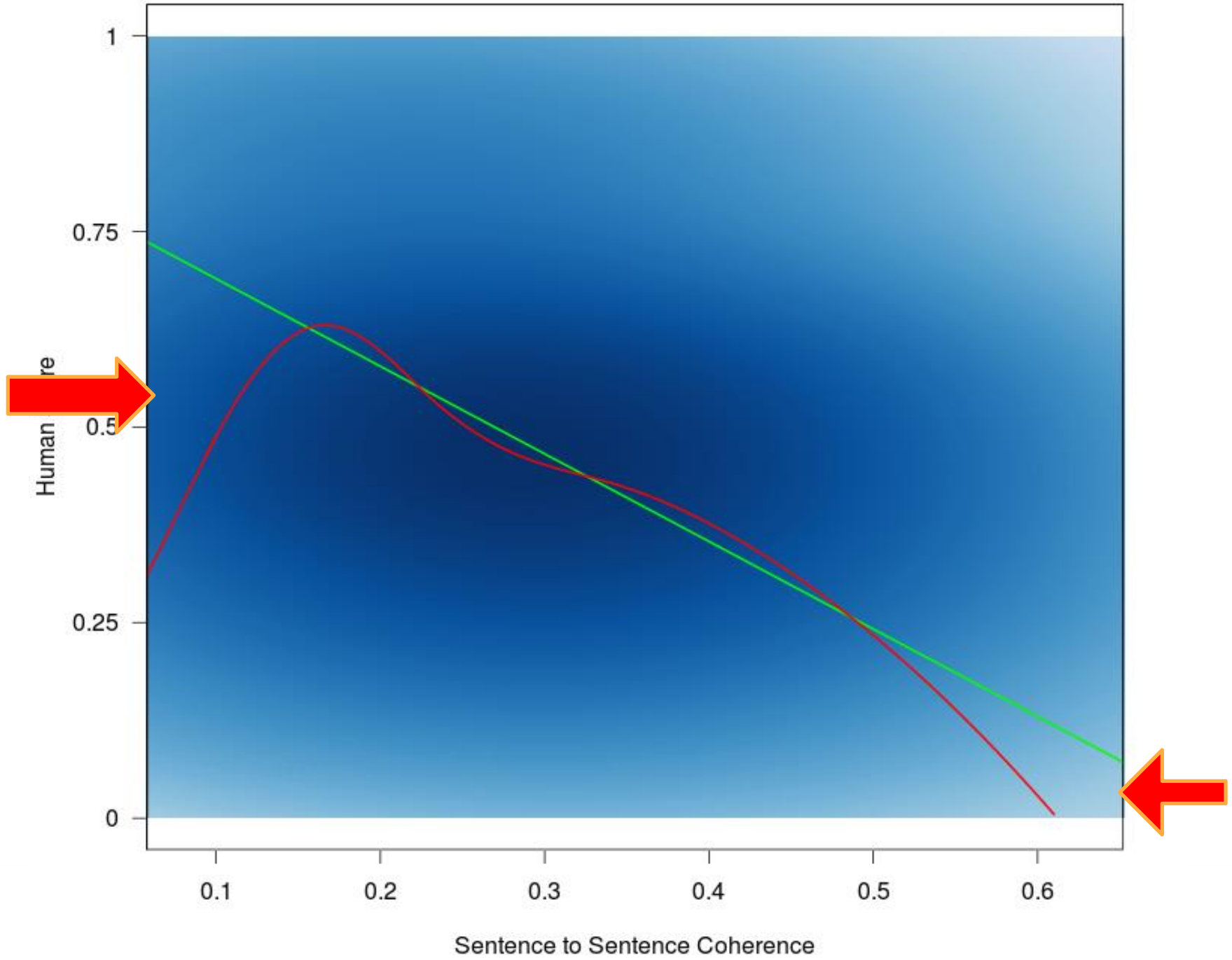
Many modeling techniques assume features behave in a linear fashion

- More (or less) is better
- Length, Content, Grammar errors, etc.

Examine features determine if linear modeling is most appropriate

Or values beyond a certain range violate the linear assumption




1
0.75
0.5
0.25
0
Human re
0.1
0.2
0.3
0.4
0.5
0.6
Sentence to Sentence Coherence

Understand the assumptions: Bias

AI scoring learns from human-based data

Does AI learn biases from humans?

SHARE

0

REPORT

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan[1,*], Joanna J. Bryson[1,2,*], Arvind Narayanan[1,*]

+ See all authors and affiliations

Science 14 Apr 2017:
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.aal4230

Peer Reviewed
see details

Article | Figures & Data | Info & Metrics | eLetters | PDF

Write an essay about a hero and describe why this person is a hero to you.

Texts	Hero
John Glenn	0.21
Martin Luther King	0.21
Jackie Robinson	0.30
President Lincoln	0.24
Fireman	0.14
Policeman	0.15

Subgroup analyses to detect bias

For each prompt evaluate the performance of IEA for various subgroups

Calculate various agreement indices (r, Kappa, Quadratic Kappa, Exact agreement) based human-human results and compare with IEA-human results

Look at standardized mean differences (SMDs) between IEA and human scores

Flag differences for any groups based on quality criteria

Measure	Threshold	Human-Machine Difference
Pearson Correlation	Less than 0.7	Greater than 0.1
Kappa	Less than 0.4	Greater than 0.1
Quadratic Weighted Kappa	Less than 0.7	Greater than 0.1
Exact Agreement	Less than 65%	Greater than 5.25%
Standardized Mean Difference	Greater than 0.15	

[2] See Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practices, 31, 2–13.

Visibility of underlying models

AI and expert human intuition

Machine learning can leverage massive amounts of data to draw inferences

By “learning” from examples, and extracting patterns, it derives rules of what features in human behavior correspond to complex tasks

Compare a students essay simultaneously to 500 other essays on 50 language features

But the resulting model is not the same as what humans consider an “explanation”

Why did Johnny get a C?

Johnny's essay word vector, weighted by its similarity to the k most similar essays in a 300 dimensional vector space summed to 2.8

and

the number of spelling errors was .3 of a standard deviation below the mean of a corpus of 5231 student essays who were in the same grade level

and

…

Human experts often also don't have access to their exact cognitive processes to cause their intuition (Ericsson & Charness, 1994, Ericsson & Simon)

But they can generate explanations (even if they are not complete) after the fact. These explanations may not always be the same as the features they used.

Making AI *decisions* visible

Extract classes of features from models to generate explanations

Organization, Conventions, Content coverage..

Keep humans in the loop: Continuous flow in summative assessment

Continuous Flow

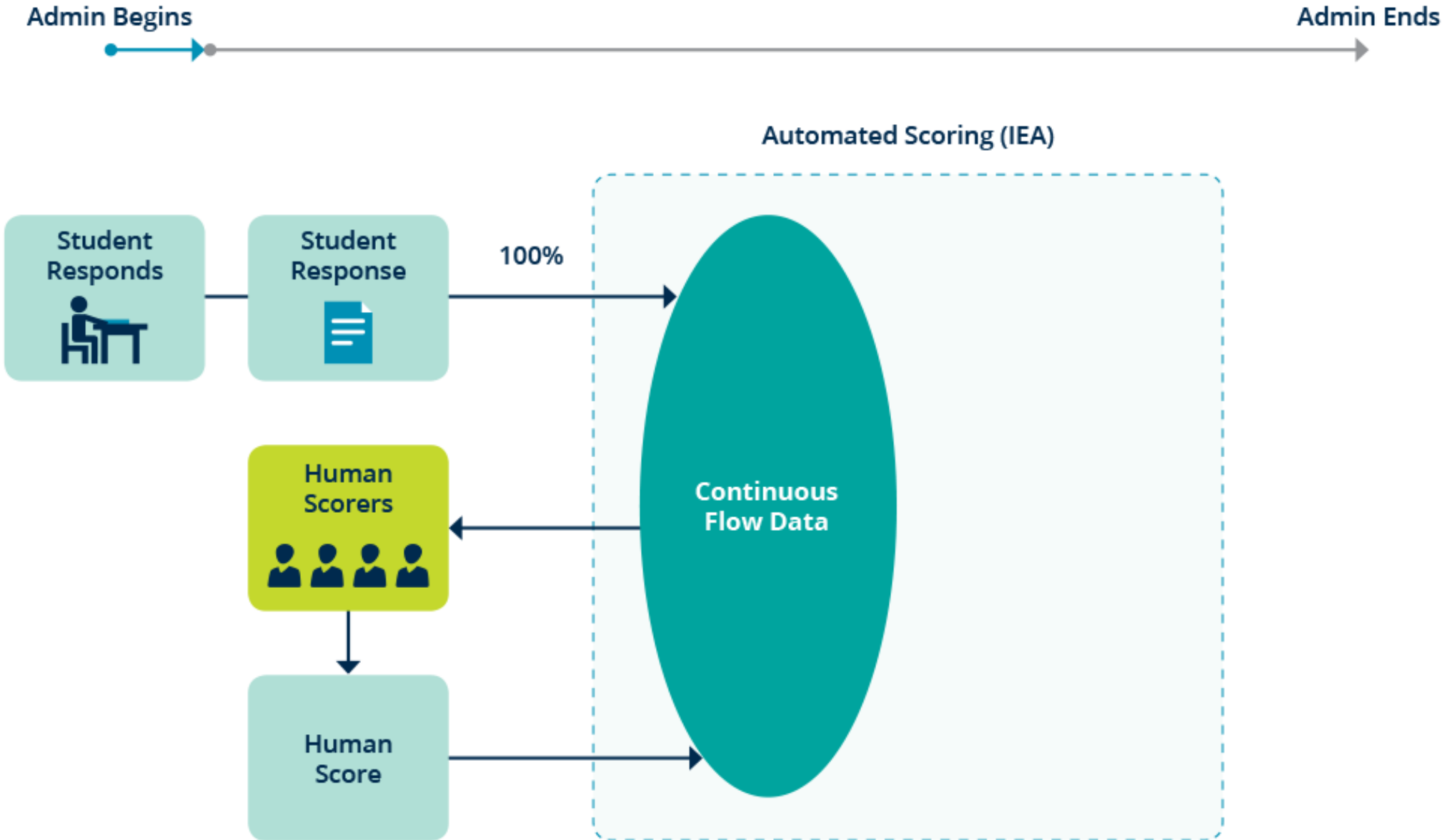
In *continuous flow* scoring, a hybrid of human and Pearson's Intelligent Essay Assessor (IEA) is used to optimize both quality and costs of scoring

Continuous flow users human scoring along with automated scoring such that responses can be branched to flow to either scoring approach

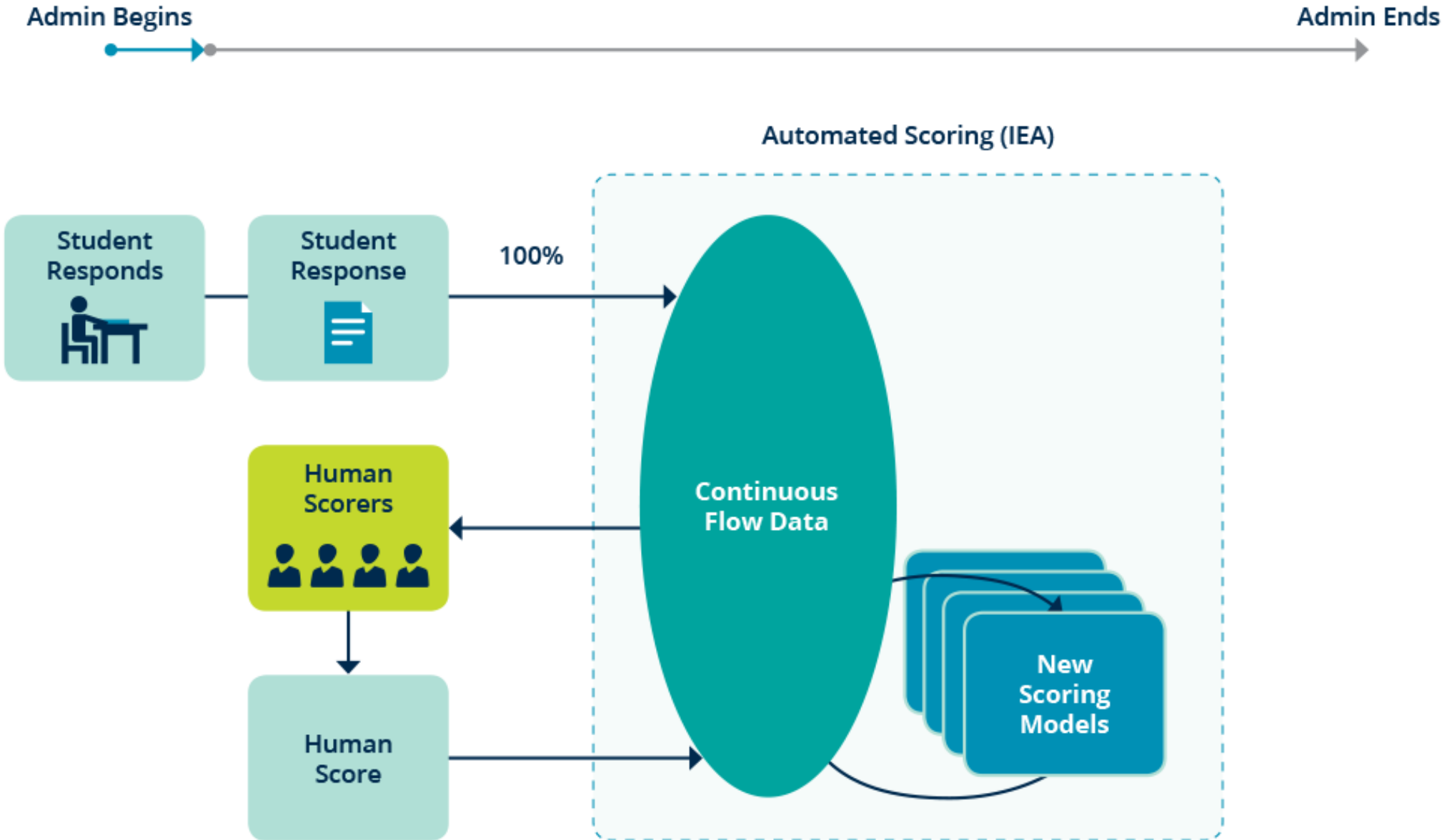
Continuous flow involves "smart routing"

> Process that automatically routes certain responses to obtain an additional human score by predicting that the automated score will be less likely to agree with humans scores.

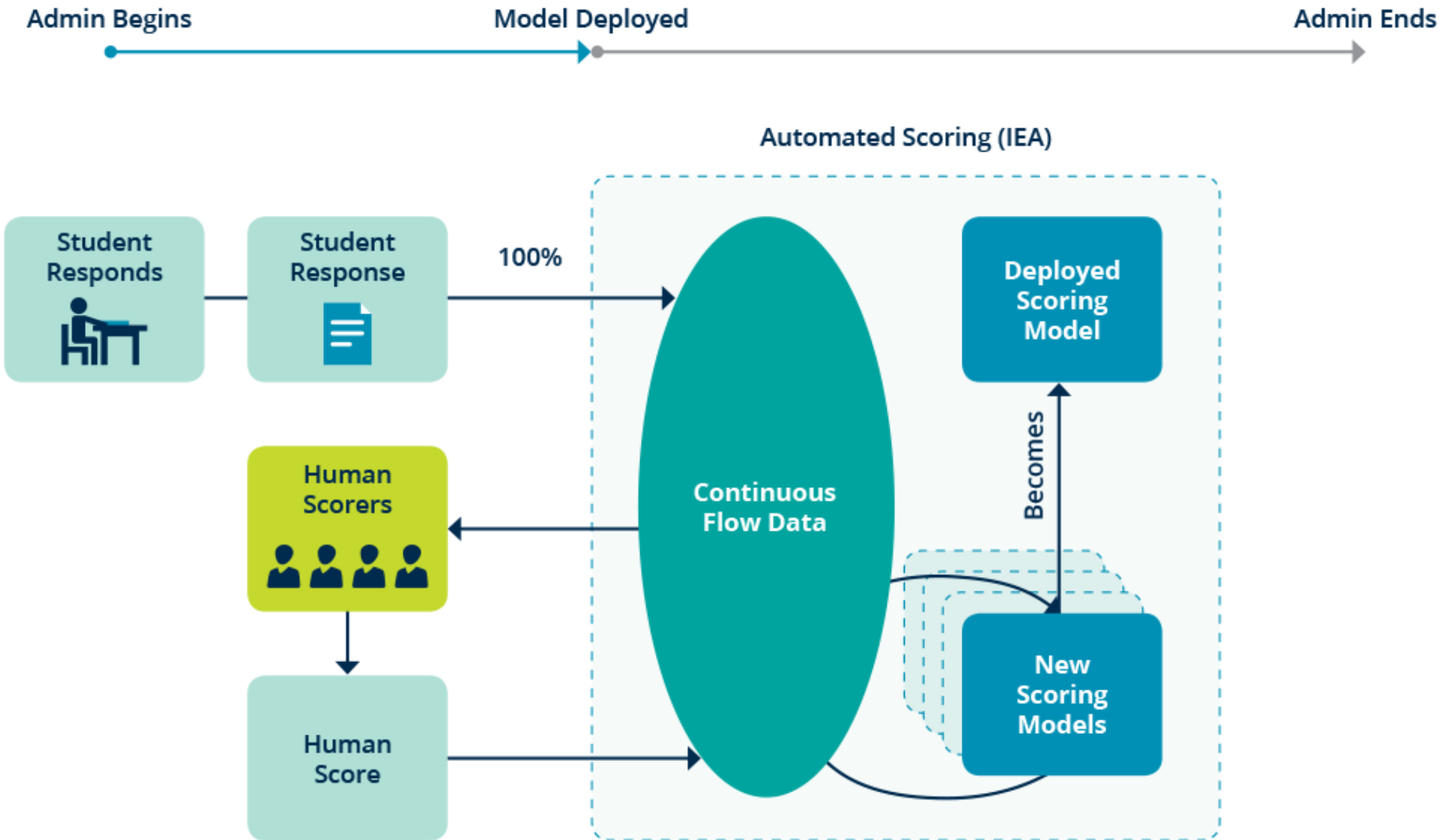
Continuous Flow Scoring: Dynamic Model Development



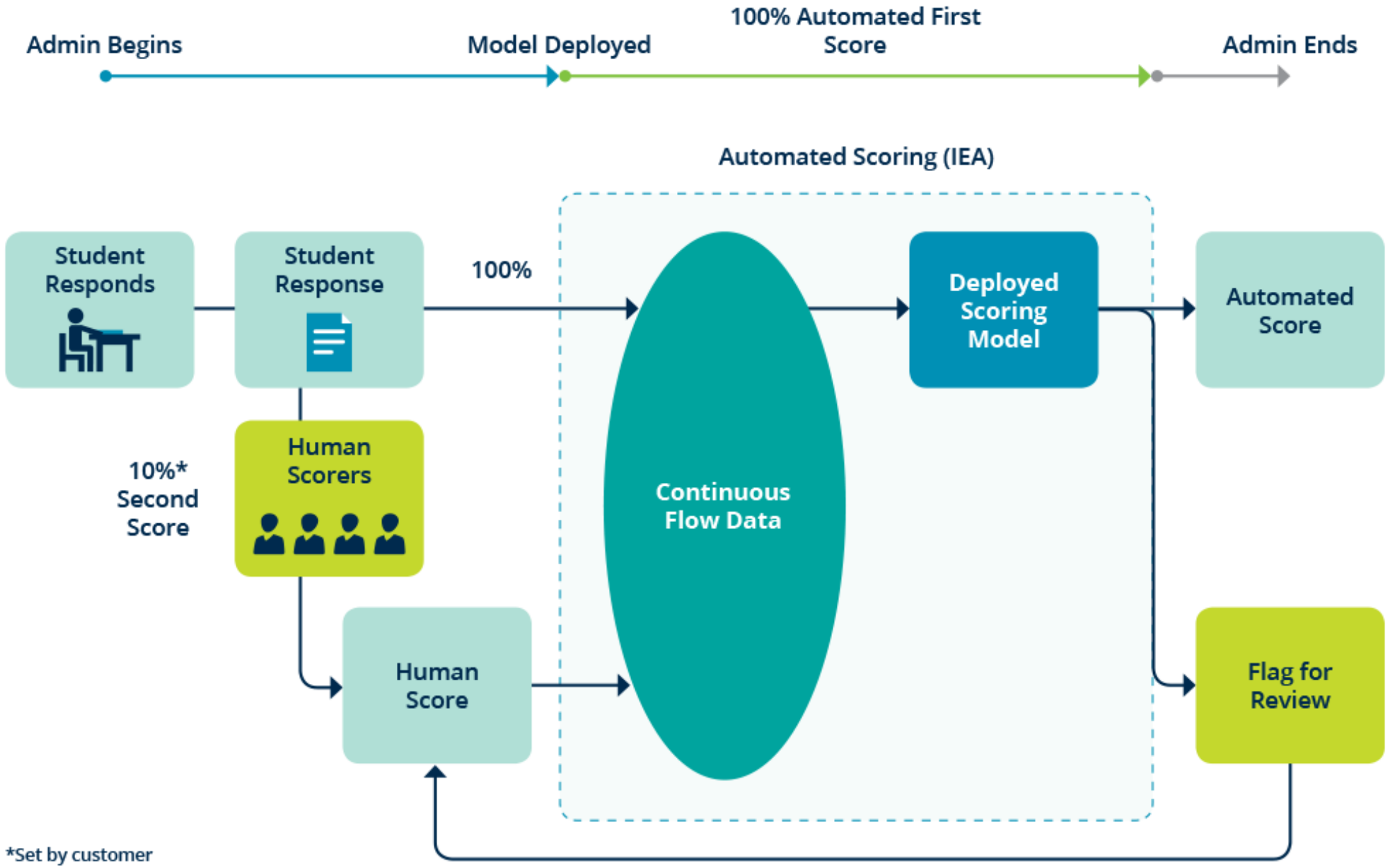
Continuous Flow Scoring: Dynamic Model Development



Continuous Flow Scoring: Dynamic Model Deployed



Continuous Flow Scoring



*Set by customer

Continuous flow conclusions

Extensive research conducted over three years to validate the use of Continuous Flow on the PARCC assessment

Successful operational use in 2016

Combines the strengths and benefits of both human and automated scoring

Performance exceeds that of a human only scoring system while routing potentially challenging responses for further review.

Keeping humans in the loop: Teachers and Students in Formative Assessment

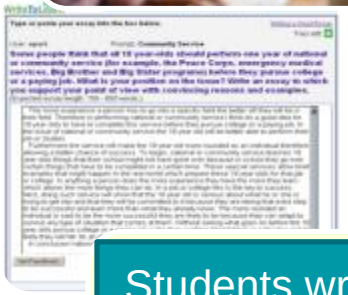
Principled Design for a Formative Writing System

Embed automated scoring in a formative system

Motivation (e.g., Graham & Perrin, 2007)

- Teaching students strategies for planning, revising, and editing their compositions (effect size=**0.82**)
- Teaching students how to assess their own writing (effect size=**0.46**)
- Explicitly and systematically teaching students how to summarize texts (effect size= **0.82**)
- Providing feedback (effect size =**0.77**)
- Monitor students' writing progress (effect size=**0.24**)

Formative writing cycle



Students write an essay to a prompt assigned by the teacher

Students get immediate feedback during writing

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

Pearson

Cameras in the Classroom -- Student Safety or Invasion of Privacy? - CCR 5

A decision has been made to equip each classroom in your school with a camera. The rationale for the placement of cameras throughout the building is student safety. Some argue this is an invasion of privacy. Think about whether you think this is a safety issue or a privacy issue. Write an essay in which you create an argument for one of these positions. Provide clear reasons and relevant evidence to support your argument.

Tips **Think about it.**

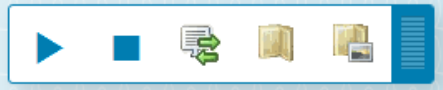
- Would you want cameras to be installed in your classroom?
- What are some of the benefits that come from security cameras in class?
- Do you think other students feel the same as you?

Draft 1

Spelling Grammar Other Tools

0 words 150 - 650 expected

Get Feedback 6 Tries Left Save



Cameras in the Classroom -- Student Safety or Invasion of Privacy? - CCR 5

A decision has been made to equip each classroom in your school with a camera. The rationale for the placement of cameras throughout the building is student safety. Some argue this is an invasion of privacy. Think about whether you think this is a safety issue or a privacy issue. Write an essay in which you create an argument for one of these positions. Provide clear reasons and relevant evidence to support your argument.

Tips **Think about it.**

- Would you want cameras to be installed in your classroom?
- What are some of the benefits that come from security cameras in class?
- Do you think other students feel the same as you?

Draft 1

Spelling Grammar Other Tools

Their are benafits to having camaras in the classroom, but it is a privacy issues. Cameras in a classrooms are creapy. Someone watching over you all the time. There's a lot of bad things at schools. I wouldnt pay attention in my class. The cameras are probubly pretty expensive and schools would spend the money on them insted of the students. That will be rong. I think it would make students fearfull also and felt picked on. The money should be spent on materials and supplies for students not on cameras. You can stop bad things from happening like fights and crime. Maybe if your iPod is stolen the cameras would catch that. that would be cool. School is supposed to be safe.

B I U

123 words 150 - 650 expected

Get Feedback Save

Cameras in the Classroom -- Student Safety or Invasion of Privacy? - CCR 5

A decision has been made to equip each classroom in your school with a camera. The rationale for the placement of cameras throughout the building is student safety. Some argue this is an invasion of privacy. Think about whether you think this is a safety issue or a privacy issue. Write an essay in which you create an argument for one of these positions. Provide clear reasons and relevant evidence to support your argument.

Tips **Think about it.**

- Would you want cameras to be installed in your classroom?
- What are some of the benefits that come from security cameras in class?
- Do you think other students feel the same as you?

Draft 2

8 Spelling | 6 Grammar | Other Tools

Their are benafits to having camaras in the classroom, but it is a privacy issues. Cameras in a classrooms are creapy. Someone watching over you all the time. There's a lot of bad things at schools. I wouldnt pay attention in my class. The cameras are probubly pretty expensive and schools would spend the money on them insted of the students. That will be rong. I think it would make students fearfull also and felt picked on. The money should be spent on materials and supplies for students not on cameras. You can stop bad things from happening like fights and crime. Maybe if your iPod is stolen the cameras would catch that. that would be cool. School is supposed to be safe.

B I U

123 words 150 - 650 expected

Get Feedback **5 Tries Left** Saved

Feedback for Draft 1

You should improve your essay. Use the bars below to see where to focus your effort.

	Passing	
Overall	2	
Task and Focus	2	Improve
Development of Ideas	1	Learn
Organization	2	Improve
Language and Style	2	Improve
Conventions	1	Learn

Good Check Repeated

Task and Focus Score: 2

English

lick to show Task

Task | Completeness | Focus

Make your claim clear.

- One part of your task is to **write a claim**.
- Your claim is a statement giving your position on the topic given in the prompt.
- In your essay, you will explain your claim with reasons and then develop those reasons with evidence.
- State your claim clearly.
  - **Clear claim:** *Students should be required to wear uniforms to school.*
  - **Unclear claim:** *I agree with this new policy.*
  - **Unclear claim:** *It's a good idea.*

Write a clear counterclaim for your essay.

- Another part of your task is to **write a counterclaim**.
- Your counterclaim is a statement about the opposing position or other side of the argument.
- In your essay, you will explain your counterclaim and tell why it isn't as reasonable as your claim.
- Write a clear counterclaim.
  - A counterclaim can give the opposite side of the issue in a general way.
    - **An adequate counterclaim that needs additional information to make it clear:** *Some people think students should not be required to wear uniforms to school.*
    - **Unclear counterclaim:** *Some people may disagree with this policy.*
  - A counterclaim can give the other side of the issue in a more specific way.
    - **A clear, specific counterclaim:** *Some parents believe that this policy will require them to spend more money on back-to-school clothing.*

Reports

Class Scoreboard | Overview | Progress | Students | Portfolio

Class: Language Arts

Activity: Cameras in the Classroom -- Student Safety or Invasion of Privacy? - CCR 5

Report Date: Thu Apr 30 2015, 10:45 am

Cameras in the Classroom -- Student Safety or Invasion of Privacy? - CCR 5

Began: Mon Apr 27 2015

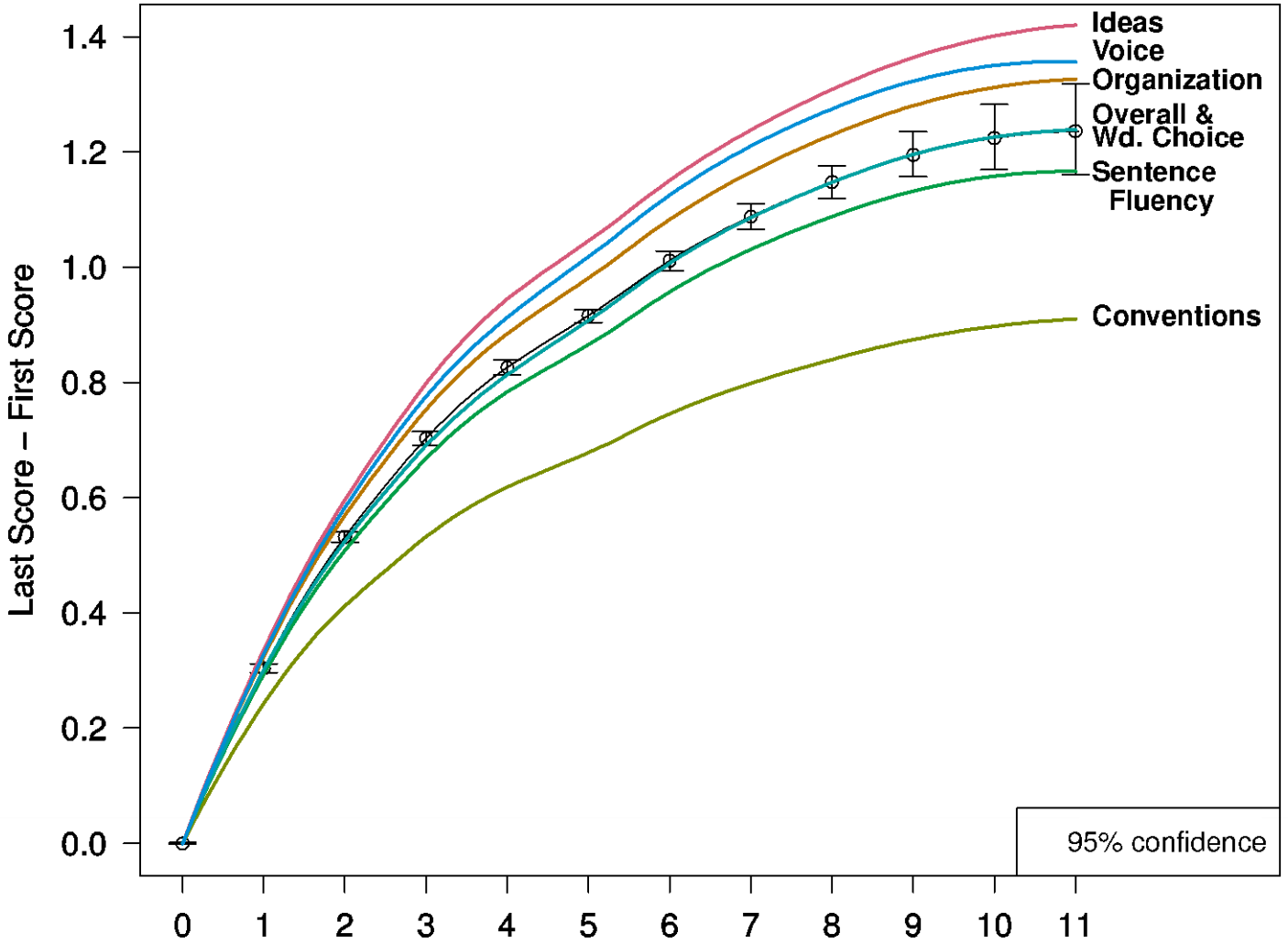
Student	Scores on Last Essay (Max. Possible)						Counts and Error Percentages						
Name (Login)	Overall (4)	Task and Focus (4)	Development of Ideas (4)	Organi-zation (4)	Language and Style (4)	Conven-tions (4)	**Word Count** (150 - 650)	**Spelling Errors**	**Grammar Errors**	**Repeated %**	**Minutes on Task** *	**Attempts**	**Passing Attempts**
Five, Student5 (sfive)	3	3	3	3	3	4	264	0	0	0	31	3	1
Four, Student4 (sfour)	2	2	2	2	2	2	146	8	6	15	24	2	0
One, Student1 (sone)	3	4	3	3	3	4	463	0	0	0	33	2	1
Six, Student6 (ssix)	2	2	1	2	2	1	123	8	6	0	12	1	0
Three, Student3 (sthree)	1	1	1	1	2	2	62	0	1	0	10	1	0
Two, Student2 (stwo)	2	2	2	2	2	3	201	4	0	7	18	1	0
Averages	Overall (4)	Task and Focus (4)	Development of Ideas (4)	Organi-zation (4)	Language and Style (4)	Conven-tions (4)	**Word Count** (150 - 650)	**Spelling Errors**	**Grammar Errors**	**Repeated %**	**Minutes on Task** *	**Attempts**	**Passing Attempts**
Students: 6 With attempts: 6	2	2	2	2	2	3	210	3	2	4	21	2	0

Download Text (CSV) Report

Jump to top

Improvement over revisions with feedback

21,137 students wrote to 72,051 assignments on 107 different unique writing prompts. 255,741 total essays



Pearson

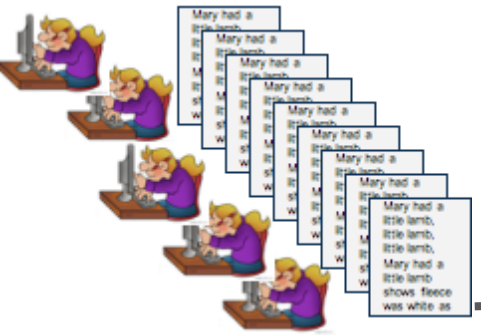
How do we automate creating scoring for formative feedback: Learning to grade like a teacher

Use real-time student writing and teacher rating to create scoring models

Allow the teacher to integrate automated scoring for their own prompts

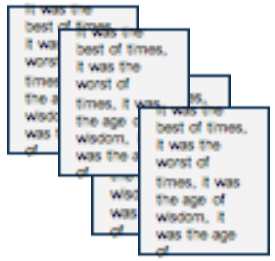
Teacher-created automated prompts with humans in the loop

Teacher assigns a writing prompt



Students write answers to prompt

System picks a subset of answers



Teacher grades the subset of answers

1
4
2
5
...

System builds automated scoring model on smaller set and grades remaining answers

1 3 4 1
3 5 6 1
3 1 3 4
4 4 6 2
3 4 3 1
...

Teacher verifies scoring of remaining answers

Scoring not OK

Scoring OK?

New prompt scoring service deployed for reuse by teacher

System selects additional essays for teacher to score

Conclusions: Lesson Learned and best practices

How I Learned to Stop Worrying and Love A.I.

By ROBERT A. BURTON date published SEPTEMBER 21, 2015 6:50 AM date updated September 21, 2015 6:50 am

23 Comments

The Stone is a forum for contemporary philosophers and other thinkers on issues both timely and timeless.

Garry Kasparov on a television monitor at the start of the final match against IBM's Deep Blue computer in May 1997 in New York.

Lessons Learned: Implementing Automated scoring into real-world large-scale contexts

Human + Computer is better than either alone: Design assessment and educational systems that leverage the best of human and machine abilities

- Computers: Fast, precise, consistent, can tell when they need a human
- Humans: More detailed, nuanced feedback, better able to connect with students

Use diverse multi-skilled teams. It creates more headaches when building solutions, but in the end, it results in better outcomes for students and educators

Best practices

Understand the assumptions that go into the modeling

- Principled design of AI techniques and features
- Consider how to make the models visible/explainable
- Evaluate how modeling may introduce bias
- Implement methods to handle edge cases

Keep humans in the loop

- Summative Scoring
- Formative Applications

As a community, we need to establish and maintain best practices and effective communication about innovations in education

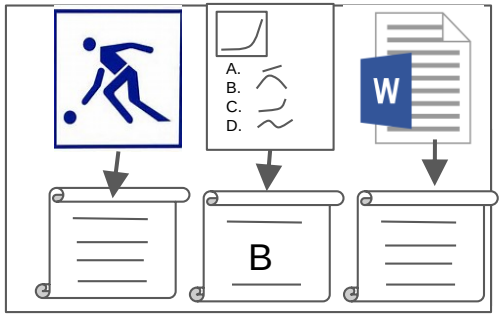
Pearson

Questions?

Peter W. Foltz

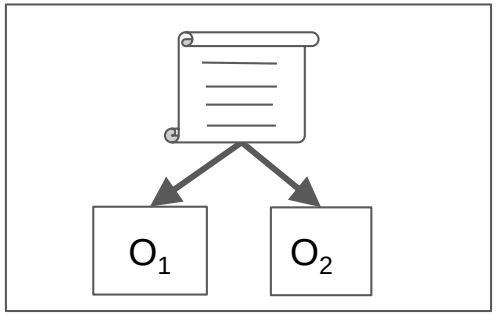
Peter.foltz@pearson.com

Task Model

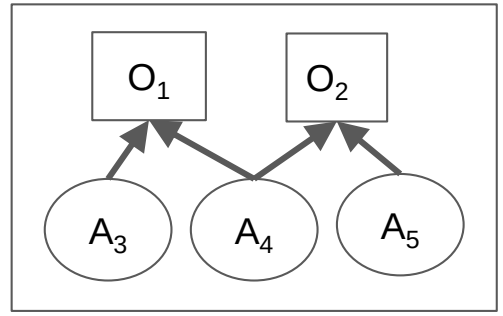


Scoring

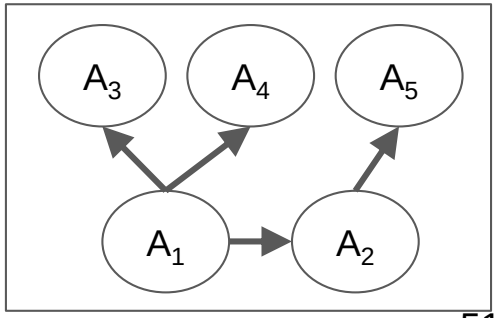
Evidence Model



Weighting

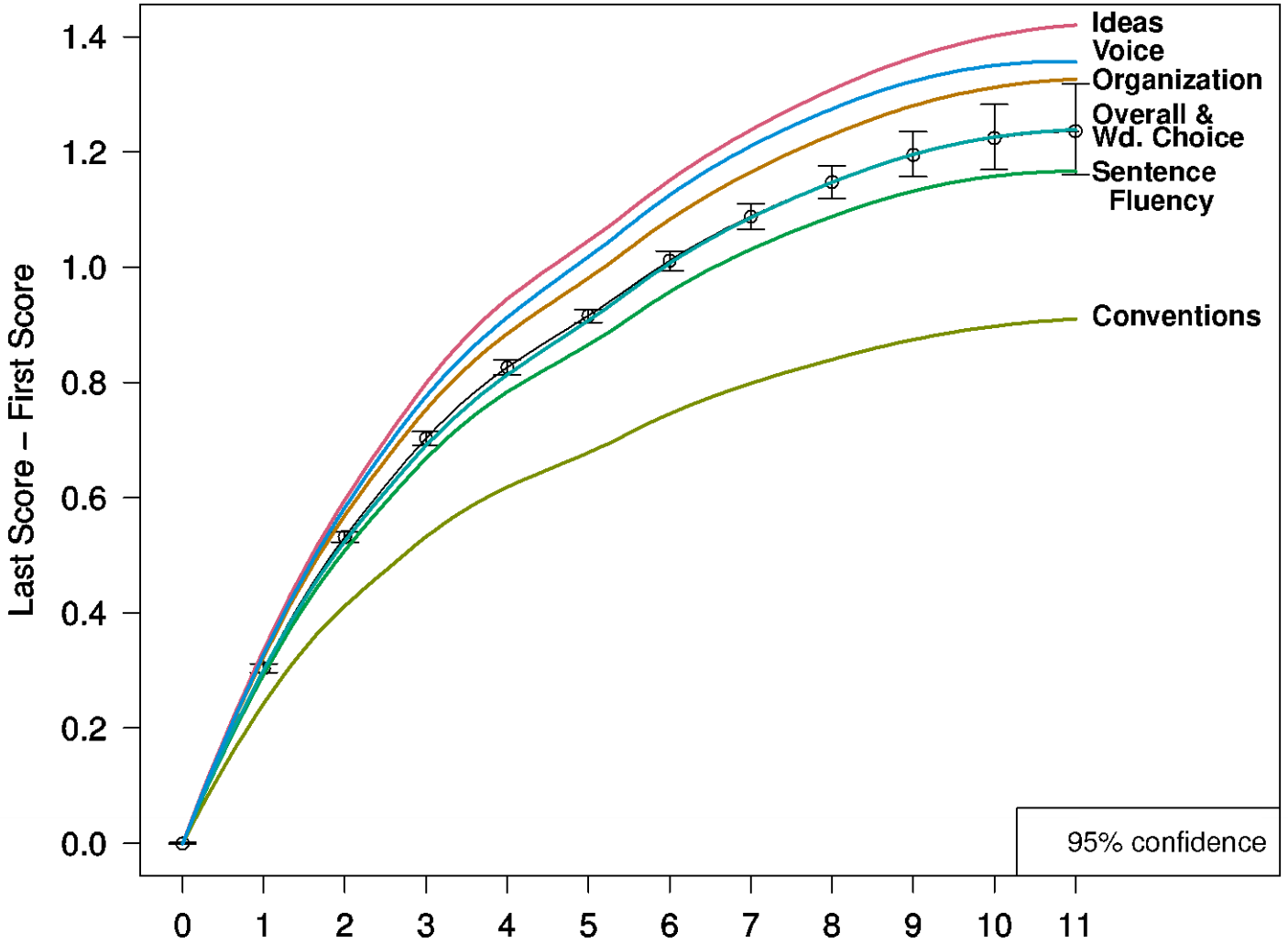


Student Model



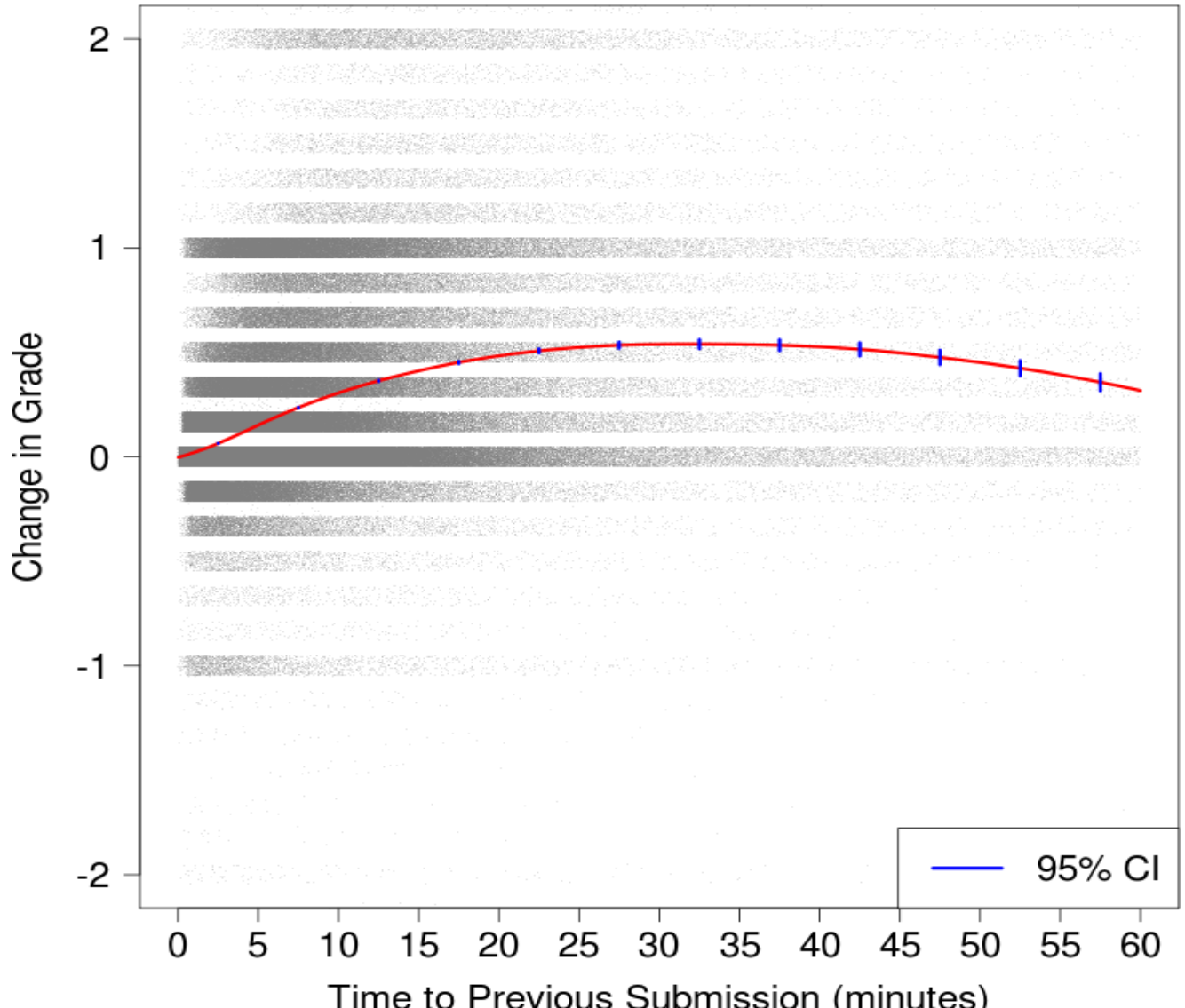
Improvement over revisions with feedback

21,137 students wrote to 72,051 assignments on 107 different unique writing prompts. 255,741 total essays



Pearson

Time between revisions: 1.1M essays



Lessons Learned: Principled Task Design for developing computer scoreable writing prompts (and writing analytics in general)

Good Prompts for automated scoring

Focused to elicit a precise range of responses

Elicits high degree of agreement between human scored

Scoring Features

Must be developed to align to constructs of interest

Not all constructs can be represented by features

Scoring rubrics

Must contain elements that are computationallly operationalizable through features

Scoring model

Should have high degree of agreement with human scorers, but matching humans is does not mean validity

Not (very) susceptible to construct irrelevant variance

Human (designer) intuition in building scoring models is not always as good as relying on machine learning and big data

Lessons Learned: Formative writing environments

Task models

You can’t control full pedagogical/task model for all environments and use cases

Fidelity of models

Balance the fidelity of scoring the writing construct to the fidelity of the activity

Less than perfect feedback is better than none

Adherence to your principles

Pedagogical purity can not always be maintained

“Sales says that you need a grammar checker”

Training IEA on Operational Data

- Continuous flow makes it relatively easy to train the automated scoring engine on operational data early in the administration window
- During this process, multiple human scores can be requested and any backreading scores assigned by supervisors can also be used to obtain the best possible data to train IEA
- Human scoring is monitored closely and when criteria are met, IEA modeling takes place
- Once IEA is trained on a particular prompt, results are evaluated by comparing IEA-human scoring agreement with human-human scoring agreement

Reporting with Multiple Scores: Best (i.e., Highest Quality) Score Wins

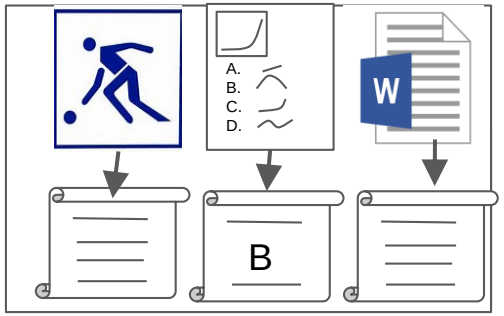
- Although multiple scores may be assigned for a given response, only one can be reported
- When multiple scores exist, there is a hierarchy for deciding which score is actually reported
  - When the automated score is the only score, it is reported
  - When there is an automated score and a human score, the human score is reported
  - When there are two human scores, the first score is reported
  - When there is a supervisor back read score, the back read score is reported
  - When there are two non-adjacent scores, a resolution score is provided and the resolution score is reported

Smart Routing Concept

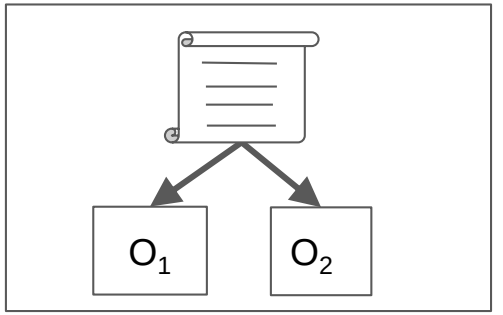
Evidence Centered Design Approach

What tasks or situations should elicit those behaviors?

Task Model



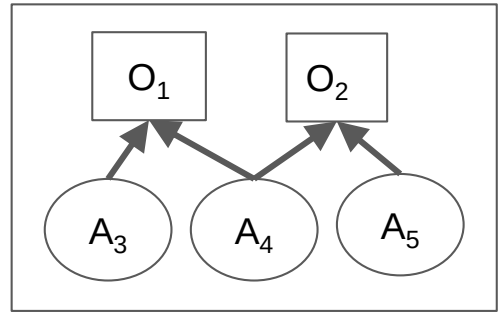
Scoring



What kind of evidence do we need to support the claims?

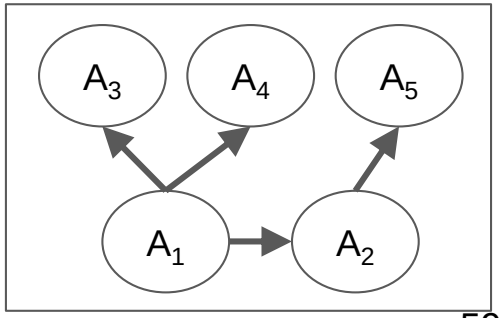
Evidence Model

Weighting

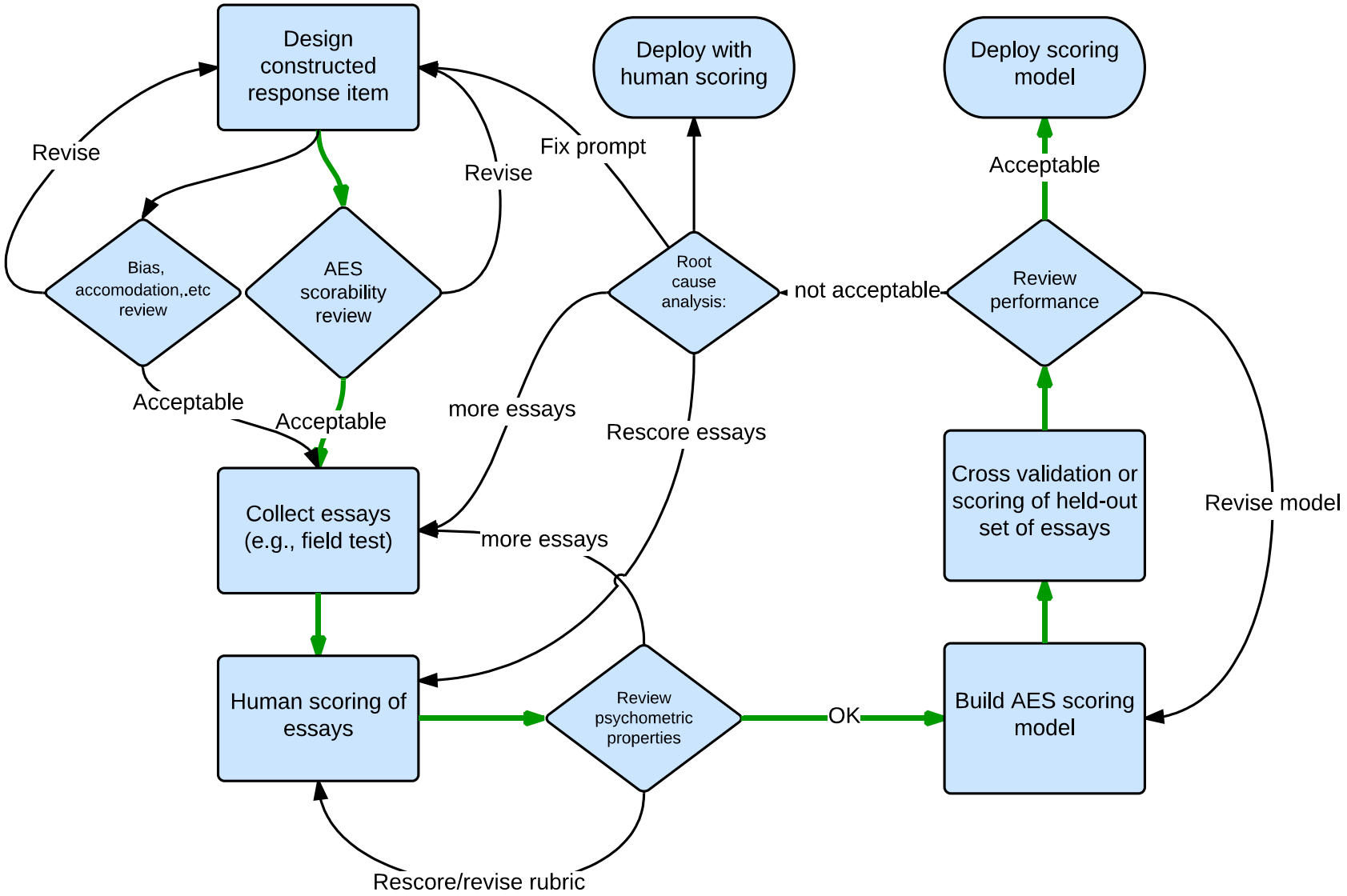


What knowledge, skills, or other attributes should be assessed and what claims made?

Student Model



High-stakes assessment workflow for automated scoring



Pearson