Using Machine Learning to Score Multi-Dimensional Science Assessments

- Sarah Maestrales
- Xiaoming Zhai
- Israel Touitou
- Quinton Baker
- Joseph Krajcik
- Barbara Schneider

Multi-Dimensional Science Assessment

National Research Council (NRC)

(NRC, 2012; NRC, 2014)

Disciplinary Core Ideas (DCI)

Crosscutting Concepts (CC)

Science and Engineering Practices (SEP)

Next Generation Science Standards (NGSS)

(NGSS, 2013)

Disciplinary Core Ideas (DCI)

- Structure of Matter
  - (3-5) Matter exists as particles that are too small to see; conservation of matter; identifying materials by their properties.
  - (6-8) Molecular composition can be used to explain properties of materials, states of matter, phase changes, and conservation of matter.

Crosscutting Concepts (CC)

- Cause and Effect
  - (6-8) Students classify relationships as causal or correlational. Can use cause and effect to predict phenomena. Sometimes phenomena may have multiple causes.
  - (9-12) Students understand that evidence is required to differentiate between cause and correlation.

Science and Engineering Practices (SEP)

- Planning and Carrying Out Investigations
  - (K-2) Simple investigations to provide evidence and support claims.
  - (3-5) Investigations that control variables.
  - (6-8) Investigations that include multiple variables.
  - (9-12) Investigations provide evidence for and test conceptual models.

Multi-Dimensional Assessment

(Harris et al.; 2019)

Automated Analysis of Student Constructed Short Answer Response

Automated Scoring

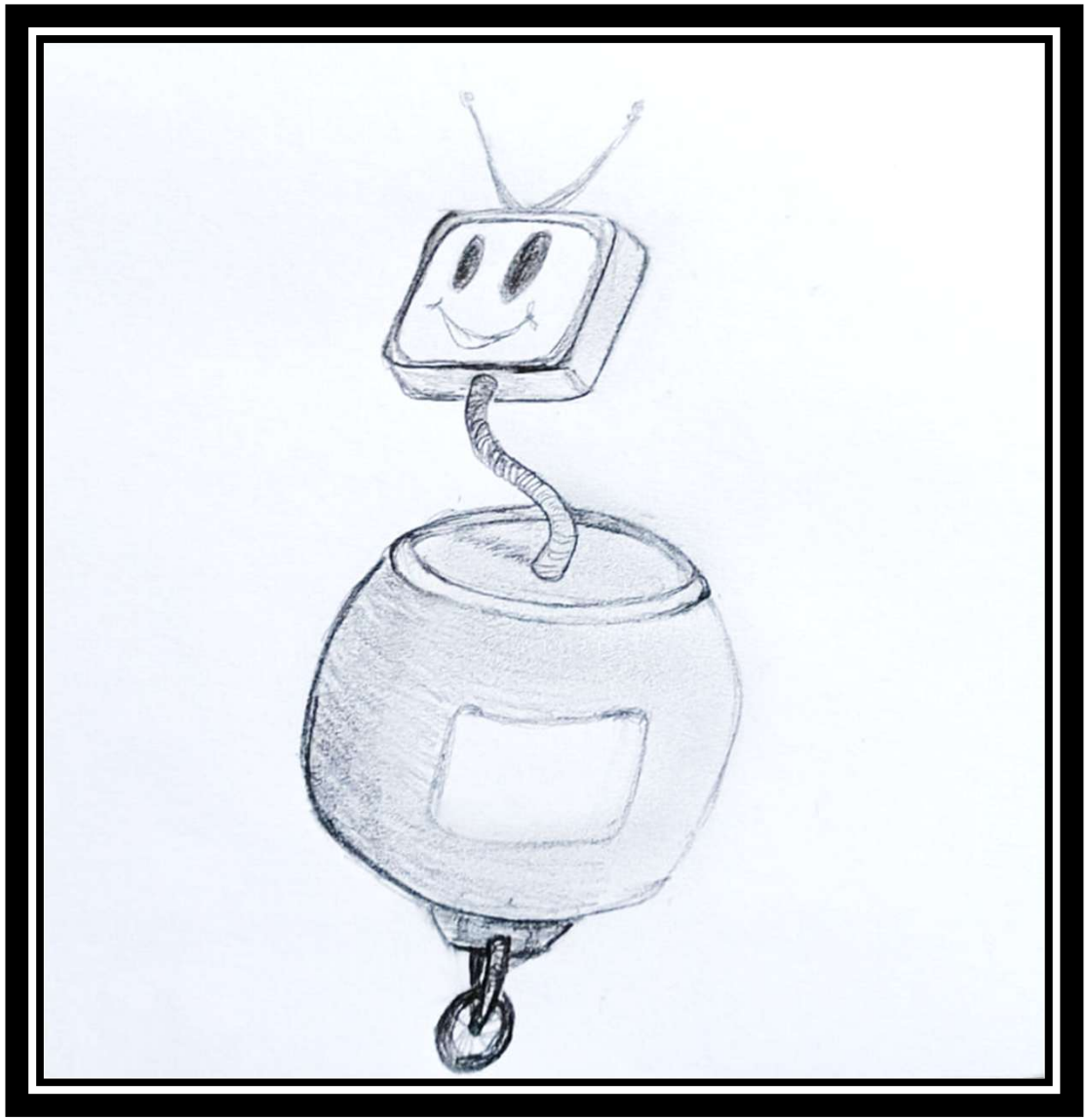
- Various researchers have used more than 20 programs or platforms to study the automatic scoring of science learning assessments (Zhai et al., 2020c).
  - A few examples:
    - Using “c-rater,” the Liu team from ETS developed multi-level rubrics and found the machine was capable of automatically scoring student responses, achieving moderate to high agreement between humans and the machine to provide information about student performance (Liu et al., 2014).
    - Mao et al. (2018) applied the c-rater-ML to provide automated feedback to students and aid them in revising their responses.
    - The earliest development of programs besides c-rater, is the SPSS Text Analysis (Nehm & Haertig, 2012), that required humans to develop word libraries manually.

Automated Analysis of Constructed Response (AACR)

- Developed to serve classroom needs for formative assessment purposes.
- Currently, it is used for both exploratory and confirmatory factor analysis in item development
- AACR Web Portal developed eight algorithms that can be employed simultaneously.

Automated Analysis of Constructed Response (AACR)

- AACR Web-Portal uses a cross validation approach by using some of the scored responses to train the algorithm while reserving the rest to test it.
- The feature extraction analysis sort the constructed responses into categories based on lexical features called “n-grams.”



Questions

1. How reliable were human raters in scoring multi-dimensional responses?
2. Could machine learning algorithms score multi-dimensional assessments as accurately as humans?
3. How are key phrases in student responses associated with machine scoring of the multi-dimensional assessments?

Methods

Crafting Engagement for Science Environments (CESE)

CRAFTING ENGAGEMENT FOR SCIENCE ENVIRONMENTS

- Conducted across two regions in the United States.
- Project based science learning intervention.
- NGSS aligned content.
- Multi-Dimensional Learning and Assessment.

OISE: 1545684

MICHIGAN STATE UNIVERSITY

CREATE for STEM Institute at Michigan State University

Sample

- 6700 High School Students
- Gender
  - 51.5% were female students
- Race:
  - 29.2% of students identified their race as white
  - 47.5% identified their race/ethnicity as Hispanic
  - 11.9% identified their race as black
  - 5.0% identified their race as Asian
- Home Language
  - 74% reported speaking Spanish in the home.

Assessment Instrument & Rubrics

- National Assessment of Educational Progress (NAEP)
  - Multiple Choice
  - Constructed Response

Assessment Instrument & Rubrics

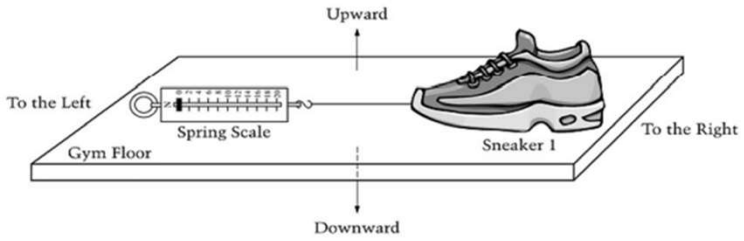
- The rubrics were created to capture whether students engaged multi-dimensional reasoning in their responses.

- While rubrics were a bit different for each item, in general the categories for classification were designed to capture the reasoning involved in the response.
  1. Incorrect – a student provides an incorrect or nonsense response.
  2. Correct - The student answered the item correctly.
  3. Multi-Dimensional Correct (MDC) – The student answered the item correctly and used reasoning associated with multiple dimensions of learning.

Item 1: Experimental Design Text and Rubric

Question text

Meg designs an experiment to see which of three types of sneakers provides the most friction. She uses the equipment listed below. 1. Sneaker 1 2. Sneaker 2 3. Sneaker 3 4. Spring scale. She uses the setup illustrated below and pulls the spring scale to the left.



Meg tests one type of sneaker on a gym floor, a second type of sneaker on a grass field, and a third type of sneaker on a cement sidewalk. Her teacher is not satisfied with the way Meg designed her experiment.

A. Describe one error in Meg's experiment.

Alignment to the NGSS (2013) Performance Expectations

Dimension	Grade-band	Performance expectation
DCI	6–8	ETS1.A Defining and Delimiting Engineering Problems
CC	(N/A)	(N/A)
SEP	3–5	Planning and Carrying Out Investigations

Rubrics

Item 1: Experimental Design Text and Rubric

Student example and multi-dimensional rubric

Multi-dimensional correct	Correct	Incorrect
"Meg's error is that she is testing three experiments in separate and different settings, allowing the experiments to have different outcomes. This stops her from knowing if her other shoes work on a gym floor or grass field or a cement sidewalk."	"Meg should have tested the sneakers in the same location for each test."	"Meg should've used different types of sneakers, not the same."
DCI: Student correctly identifies the error in the experimental setup. & SEP: Student explains this is a failure to control for variables or that the results cannot be compared.	DCI: Student correctly identifies an error in the experimental setup. & No SEP: Student does not explain that it controls for relevant variables.	Provides an incorrect response or irrelevant error in the experimental setup.

Item 2: Relative Motion Text and Rubric

Question text

Suppose you are riding in a car along the highway at 55 miles per hour when a truck pulls up along the side of your car. This truck seems to stand still for a moment, and then it seems to be moving backward.

A. Tell how the truck can look as if it is standing still when it is really moving forward.

Alignment to the NGSS (2013) Performance Expectations

Dimension	Grade-band	Performance expectation
DCI	6–8	PS2.A Forces and Motion
CC	6–8	Scale and Proportion
SEP	(N/A)	(N/A)

Rubrics

Item 2: Relative Motion Text and Rubric

Student example and multi-dimensional rubric

Multi-dimensional correct	Correct	Incorrect
"The truck looks as if it is standing still as both your car and the truck are moving at 55 mph in the same direction."	"It is going 55 miles per hour, which is as fast as the car is going."	"the truck looks like it is still because it is losing speed."
DCI: Student relates the truck's speed to the speed of the observer. & CC: Student states that equal relative speeds would cause the truck to appear as though it is standing still.	DCI: Student relates the truck's speed to the speed of the observer. & No CC: Student does not discuss the visual phenomenon being caused by the relative speeds.	Student provides an incorrect/irrelevant explanation for the phenomena OR only restates the question.

Item 3: Properties of Solutions Text and Rubrics

Question text

Maria has one glass of pure water and one glass of salt water, which look exactly alike. Explain what Maria could do, without tasting the water, to find out which glass contains the salt water.

Alignment to the NGSS (2013) Performance Expectations

Dimension	Grade-band	Performance expectation
DCI	3–5, 6–8	PS1.A Structure and Properties of Matter
CC	6–8	Cause and effect
SEP	6–8	Planning and Carrying Out Investigations

Rubrics

Item 3: Properties of Solutions Text and Rubrics

Student example and multi-dimensional rubric

Multi-dimensional correct	Correct	Incorrect
"Maria could use two similar cups and weigh them both and the heavier one is saltwater."	"Maria can weigh the cups that hold the water."	"Your body floats easier in salt water."
SEP: Student response describes an experiment that controls for relevant variables.	SEP: Student response describes an experiment that controls for relevant variables.	Student response does not describe an experiment that will differentiate fresh water from salt water.
DCI: The experiment isolates a measurement that will differentiate fresh water from salt water.	DCI: The experiment isolates a measurement that will differentiate fresh water from salt water.	
CC: Student indicates the expected result that will allow them to differentiate the fresh water and salt water.	No CC: Student does not indicate the expected result that will allow them to differentiate the fresh water and salt water.	

Item 4: States of Matter Text and Rubrics

Question text

Anita puts the same amount of water in two pots of the same size and type. She places one pot of water on the counter and one pot of water on a hot stove.

After ten minutes, Anita observes that there is less water in the pot on the hot stove than in the pot on the counter, as shown below.

Pot of Water on Counter Pot of Water on Hot Stove

A. Why is there less water in the pot on the hot stove?

B. Where did the water go?

Alignment to the NGSS (2013) Performance Expectations

Dimension	Grade-band	Performance expectation
DCI	6–8	PS1.A Structure and Properties of Matter
CC	6–8	Energy and Matter
SEP	(N/A)	(N/A)

Rubrics

Item 4: States of Matter Text and Rubrics

Student example and multi-dimensional rubric

Multi-dimensional correct	Correct	Incorrect
"The heat caused it to evaporate."	"The water evaporated."	"it dried up."
DCI: Student says the water evaporated.	DCI: Student says the water evaporated.	Provides an incorrect/ irrelevant explana- tion.
&	OR	
CC: Attributes this to the heat from the stove.	CC: Attributes this to the heat from the stove.	

The Process

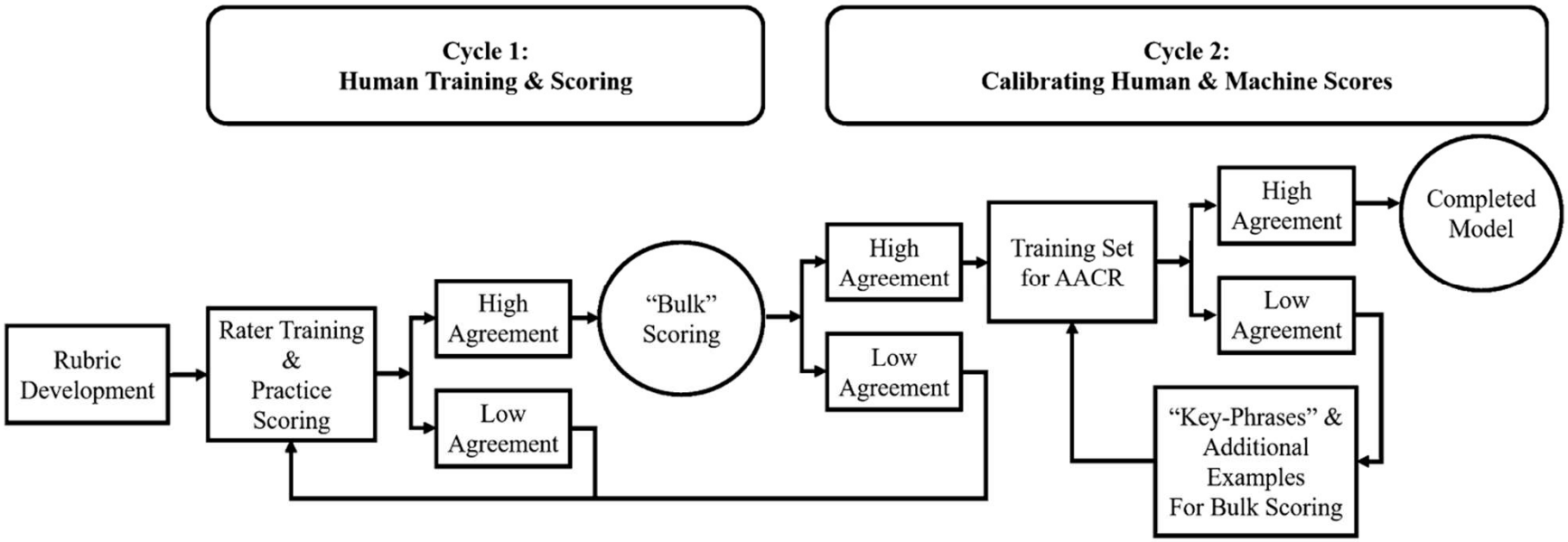


Fig. 2 Training and algorithm development

Cycle 1:
Human Scoring

Cohen's Kappa (k)

- Agreement between raters adjusted for chance agreement or guessing
- Used for categorical data

The formula to calculate Cohen's kappa for two raters is:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

where:

P_o = the relative observed agreement among raters.

P_e = the hypothetical probability of chance agreement

The Kappa statistic varies from 0 to 1, where.

- 0 = agreement equivalent to chance.
- 0.1 – 0.20 = slight agreement.
- 0.21 – 0.40 = fair agreement.
- 0.41 – 0.60 = moderate agreement.
- 0.61 – 0.80 = substantial agreement.
- 0.81 – 0.99 = near perfect agreement
- 1 = perfect agreement.

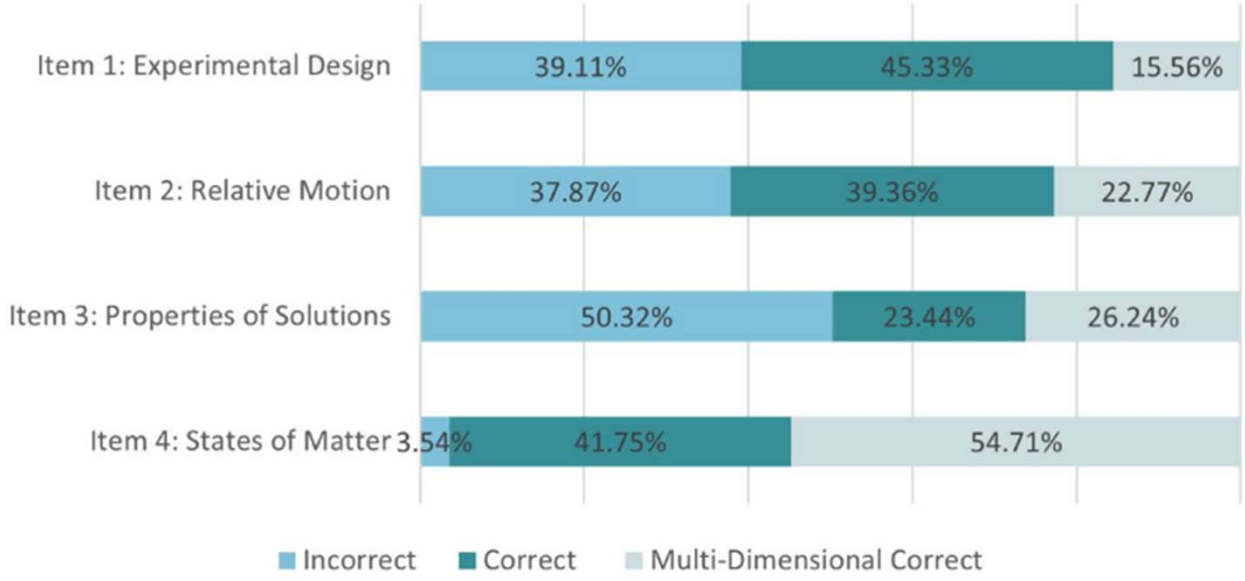
How reliable were human raters in scoring multi-dimensional responses?

- Human agreement by wave

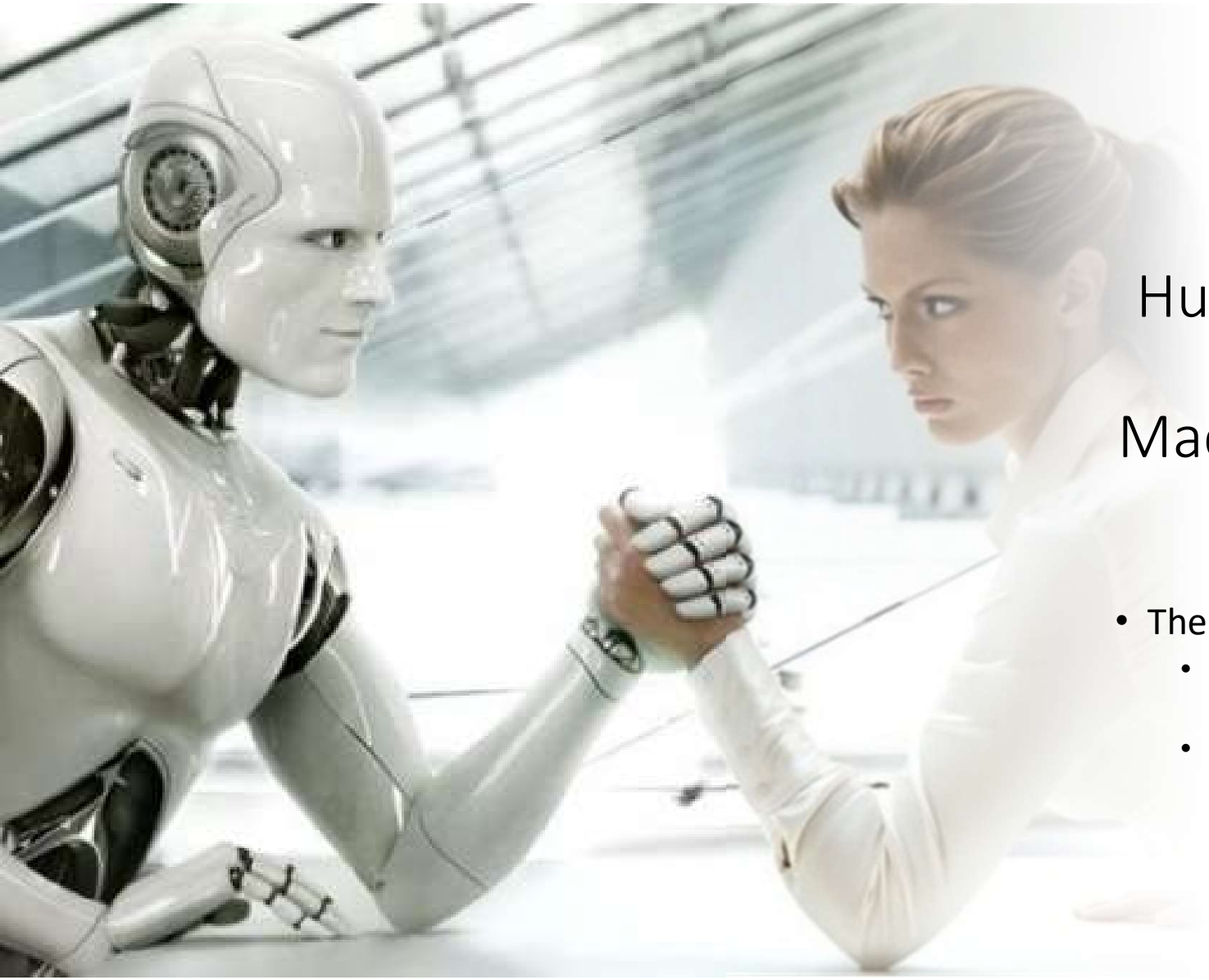
		Wave			Cumulative			
Wave	$N_{Overlap}$	Accuracy	k	$N_{Cummulative}$	Accuracy	k	N_{Total}	Note
Item 1: experimental design								
1	99	0.81	0.71	99	0.81	0.71	0	Practice
2	50	0.58	0.38	50	0.58	0.38	0	Practice
3*	100	0.84	0.73	100	0.84	0.73	0	Practice
4	25	0.80	0.67	25	0.80	0.67	0	Practice
Item 2: relative motion								
1	60	0.83	0.72	60	0.83	0.72	51	Practice
2	60	0.85	0.77	90	0.83	0.74	80	Practice
3	90	0.93	0.88	180	0.88	0.81	439	Bulk scoring
4	30	0.87	0.80	210	0.88	0.81	564	Key phrases
5	0			210	0.88	0.81	602	Key phrases
6*	33	0.85	0.76	243	0.87	0.80	740	Bulk scoring
7*	75	0.87	0.80	318	0.87	0.80	808	Bulk scoring
Item 3: properties of solutions								
1	30	0.77	0.63	30	0.77	0.63	0	Practice
2*	190	0.73	0.56	190	0.73	0.56	0	Practice
3**	70	0.72	0.57	70	0.72	0.57	70	Bulk scoring
4**	400	0.78	0.65	470	0.77	0.64	465	Bulk scoring
Item 4: states of matter								
1	30	0.93	0.86	30	0.93	0.86	88	Bulk scoring
2	19	0.79	0.62	49	0.88	0.77	242	Bulk scoring
3*	25	0.85	0.72	74	0.87	0.75	524	Bulk scoring
4*	25	0.89	0.80	99	0.87	0.76	594	Bulk scoring

$N_{Overlap}$ is the number of items raters scored in common. $N_{Cumulative}$ is the total number of jointly scored responses in all combined waves. N_{Total} is the number of total responses scored which can be sent to the machine. A wave number followed by * indicates that this wave, and those following, were scored by a new team, while ** indicates a third rater was added to tiebreak

Students' Use of Multi-Dimensional Reasoning



Cycle 2:
Machine Scoring



Human-to-Human vs Machine-to-Human Agreement

- The machine algorithm
  - Agreed well with human scores.
  - Scored somewhat lower than humans on most items.

Descriptions of human and machine Scores

Wave (sample)	Mean (*SE*)		k (*SE*)
	Human	Machine	
Item 2: relative motion			
1 (484)	1.83 (0.03)	1.76 (0.03)	0.78 (0.02)
2 (662)	1.89 (0.03)	1.83 (0.03)	0.78 (0.02)
3 (808)	1.85 (0.03)	1.81 (0.03)	0.81 (0.02)
Item 3: properties of solutions			
1 (468)	1.76 (0.04)	1.68 (0.04)	0.69 (0.03)
Item 4: states of matter			
1 (336)	2.59 (0.03)	2.60 (0.03)	0.76 (0.04)
2 (594)	2.51(0.02)	2.49 (0.02)	0.64 (0.03)

Item 1 is not shown because the item did not proceed to machine scoring

Accuracy by Dimension

- The algorithms scored well regardless of the dimensions being measured.
- The algorithms scored most accurately when
  - Each category was well represented.
  - Human raters were in high agreement.

- Human and machine percentage of score, agreement, certainty and dimensionality

Proficiency Level	Item dimensionality DCI, CC, SEP	Human	Machine	Mean Probability (SE)	Accuracy
Item 2: relative motion ($N=808$)					
Incorrect	Incorrect	37.87	39.98	90 (0.00)	93.14
Correct	DCI	39.36	38.74	91 (0.01)	86.79
MDC	DCI + CC	22.77	21.29	82 (0.01)	78.80
Item 3: properties of solutions ($N=468$)					
Incorrect	Incorrect or DCI only	50.32	57.63	85 (0.01)	92.31
Correct	DCI + SEP	23.44	16.56	78 (0.01)	59.63
MDC	DCI + CC + SEP	26.24	25.81	80 (0.01)	79.51
Item 4: states of matter ($N=594$)					
Incorrect	Incorrect	3.54	2.19	68 (0.03)	28.57
Correct	DCI or CC	41.75	46.30	83 (0.01)	83.87
MDC	DCI + CC	54.71	51.52	87 (0.01)	82.77

Item 1 is not shown because it did not proceed to machine scoring. Mean probability refers to the prediction returned from AACR that a given score was correct.

MDC multidimensional correct, *DCI* disciplinary core ideas, *CC* crosscutting concepts, *SEP* science and engineering practice

Probability of a Correct Response & Language Use

- While some phrases were over-represented among the discrepancies between human and machine scores, the AACR algorithm largely showed well represented language choices were scored incorrectly in similar proportions to their appearance in the population.

- Key phrases associate with the machine scoring

Key phrase(s)	% of all responses	% Scored incorrectly	% All MH disagreements	Mean probability (SD)
Item 2: relative motion ($N_{Responses} = 808$ and $N_{Disagreements} = 102$)				
Velocity and relative	0.50	50.00	1.96	77 (0.10)
Relative	0.87	71.43	4.90	81 (0.10)
Velocity or relative without speed	7.18	20.69	11.76	84 (0.09)
Fast	7.80	12.70	7.84	87 (0.11)
Velocity	8.79	18.31	12.75	85 (0.09)
Speed	56.06	11.70	51.96	89 (0.10)
Item 3: properties of solutions ($N_{Responses} = 465$ and $N_{Disagreements} = 87$)				
Taste	1.08	0.00	0.00	80 (0.07)
Freeze	2.15	10.00	1.15	83 (0.11)
pH	2.80	7.69	1.15	80 (0.08)
Dissolve	3.01	7.14	1.15	80 (0.13)
Smell	6.02	14.29	4.60	83 (0.10)
Mass or weight	7.74	16.67	6.90	83 (0.09)
Evaporate	7.96	21.62	9.20	83 (0.11)
Boil	10.97	19.61	11.49	82 (0.11)
Density	13.98	13.85	10.34	83 (0.10)
Item 4: states of matter ($N_{Responses} = 594$ and $N_{Disagreements} = 111$)				
Steam	37.54	16.14	32.43	86 (0.10)
Into the air	15.66	16.13	13.51	88 (0.09)
Heat and evaporation	20.37	14.05	15.32	89 (0.09)
Heat	21.21	15.08	17.12	89 (0.10)
Evaporation	37.54	15.25	30.63	86 (0.10)

Item 1 is not shown because it did not move to machine scoring. Mean Probability refers to the prediction made by AACR that the algorithm assigned the same classification as the human raters


CAUTION

Item Response Theory

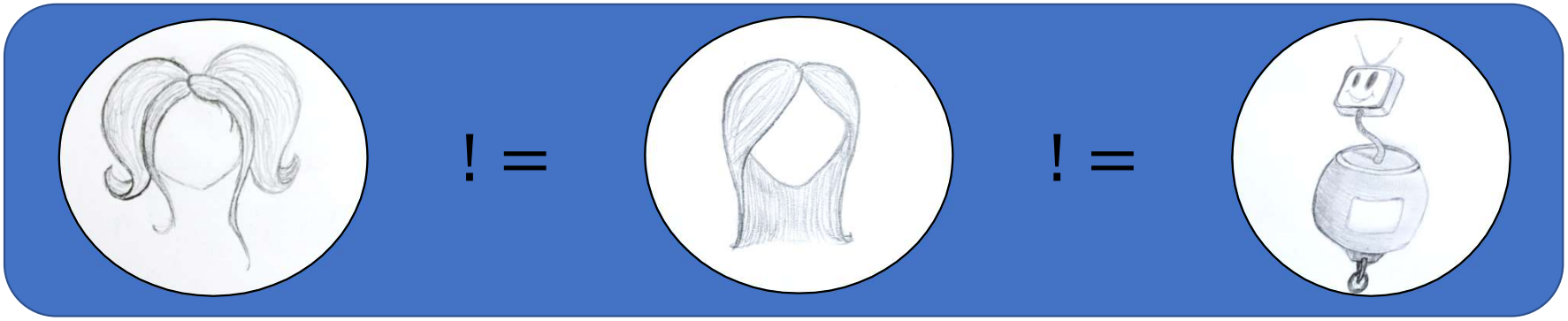
- Difficulty shows the ability level where the probability of a correct (or given) response becomes greater than 0.5 (Raykov & Marcoulides, 2018).
- Discrimination shows how well an item differentiates between students of a given level of ability (Raykov & Marcoulides, 2018).

Item parameters may differ depending upon characteristics of the individual rater (Wu, 2017)

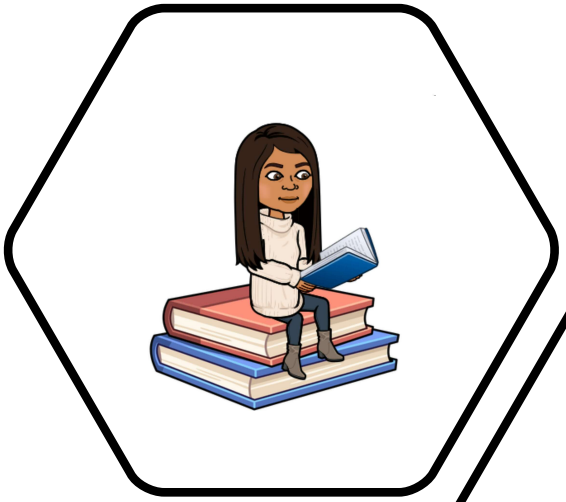
- Some raters may score higher or lower.
- Some raters may make more "mistakes" than others.

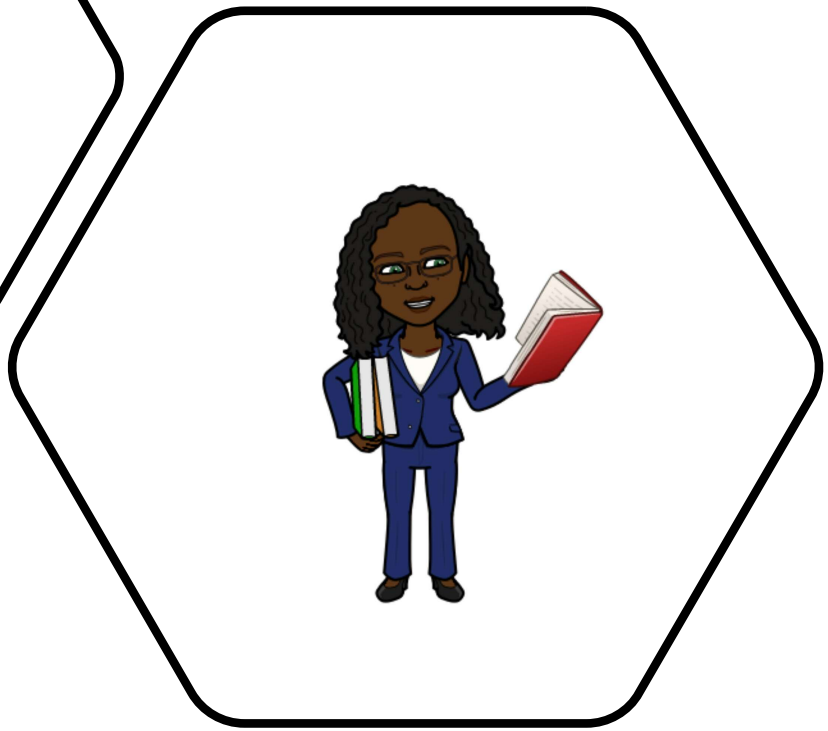
Estimates of a student's ability may differ based on which scores are used to estimate parameters (Wang & Yao, 2013; Wang & Sun, 2018).

- e.g. estimates of a given student's ability may appear lower if a rater scores lower.

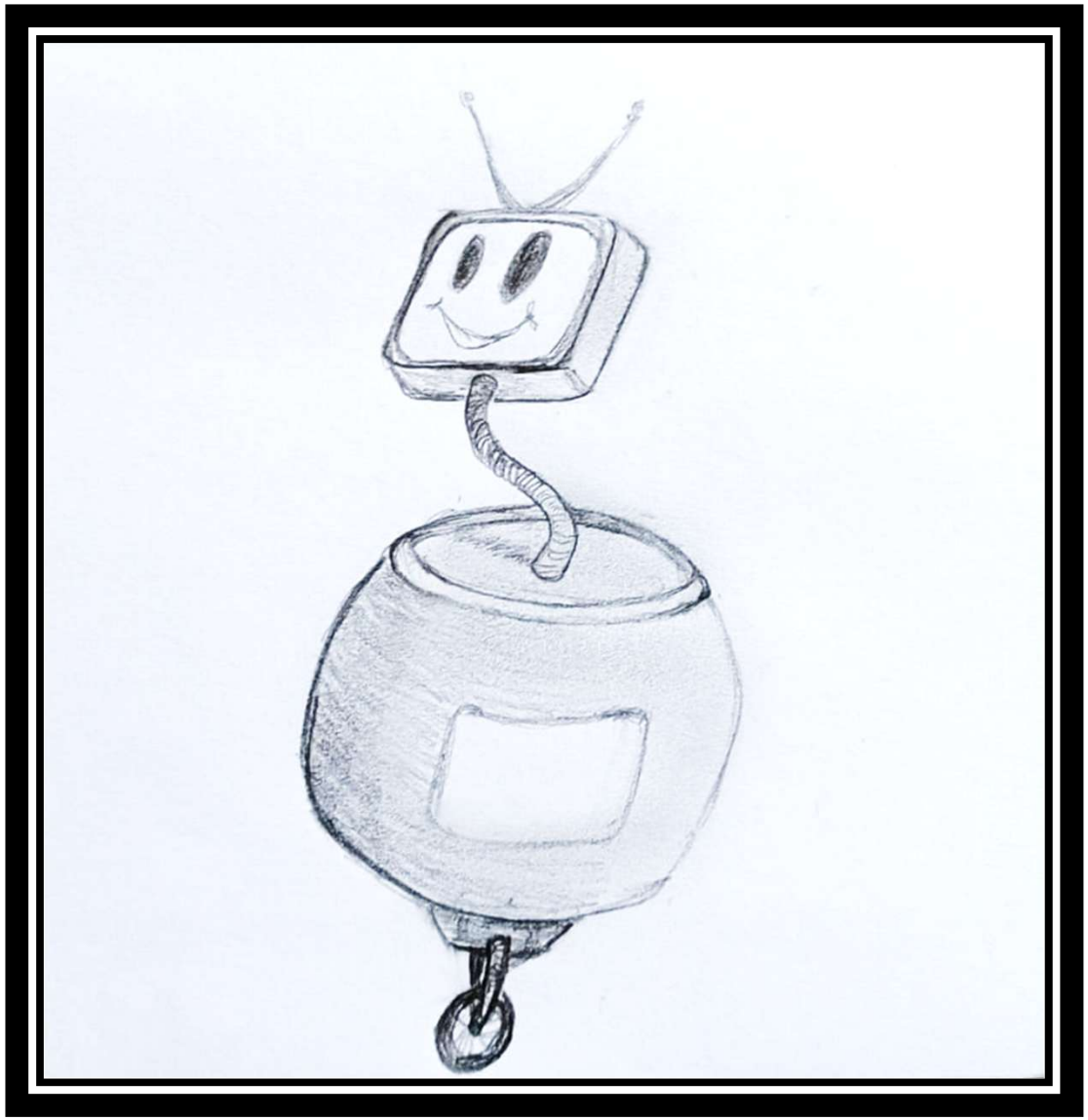


Ethical Concerns





- Some argue that machine learning algorithms may reflect bias (conscious or unconscious) of the developers and have the potential to harm minority groups, or others who are members of already disadvantaged groups (Lee, 2018; Yapo & Weiss, 2018).
- in the case of scoring academic assessments, bias could directly impact students' educational outcomes.
- Machine learning has a difficult time scoring more unique texts (Balfour, 2013), which could introduce bias if there are significant differences in language use between groups.



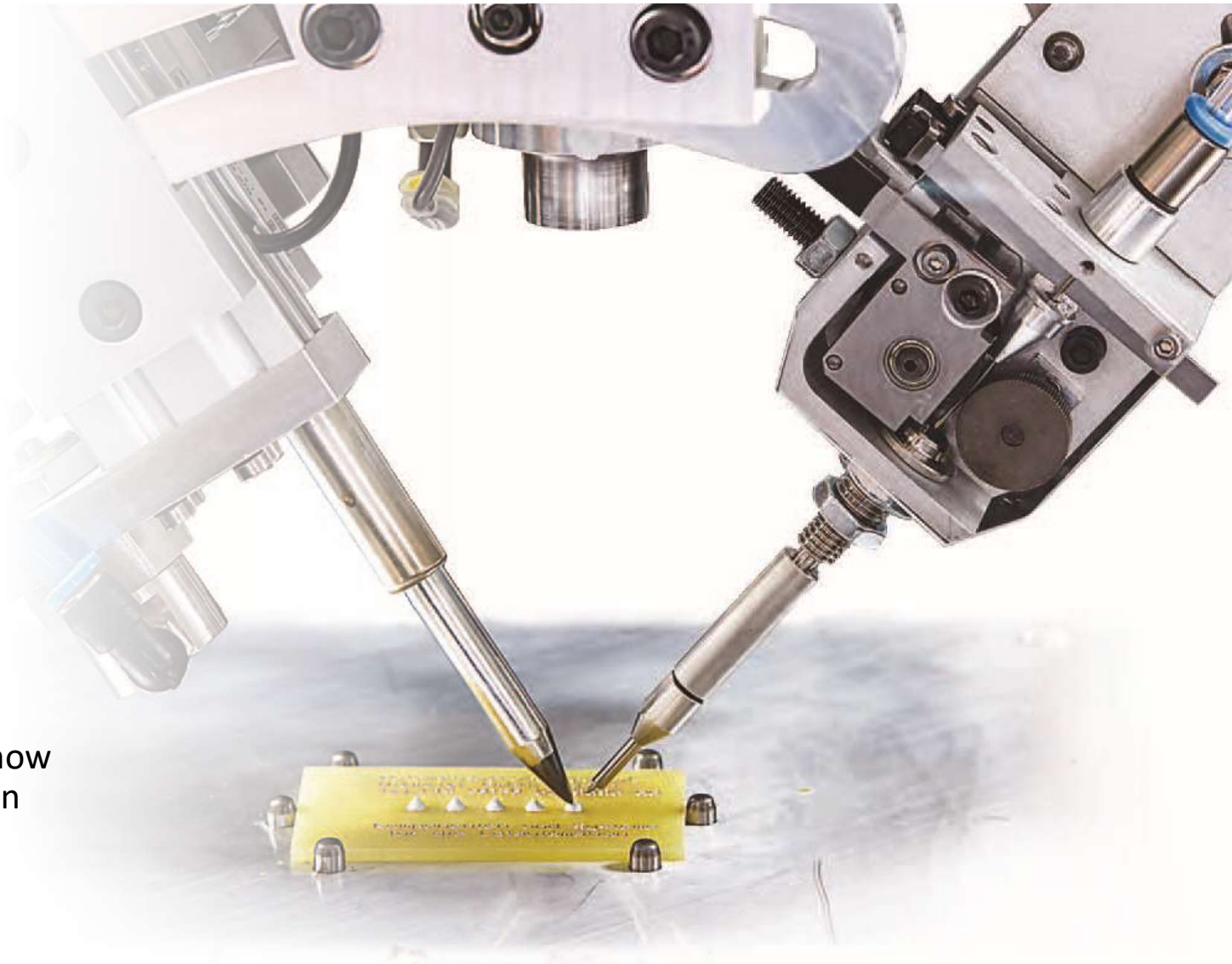
Questions

1. Do item parameters change when calculated using machine scores?
2. Is the machine algorithm likely to score differently than humans by race, ethnicity, gender, or languages spoken in the home?

Difficulty & Discrimination

There were overlaps in the confidence intervals for all difficulty and discrimination parameters.

The machine also tended to show somewhat lower discrimination than did human raters.



Difficulty and Discrimination Parameters Using Human and Machine Scores

Difficulty and Discrimination Parameters in Human vs. Machine Scoring

	Human		Machine	
	Coefficient	95% CI	Coefficient	95% CI
Item 2: Relative Motion				
Difficulty: Incorrect vs DCI	-0.27	-0.47 - -0.07	-0.14	-0.37 - 0.09
Difficulty: DCI vs MDC	1.04	0.80 - 1.29	1.21	0.91 - 1.51
Discrimination	0.91	0.68 - 1.15	0.76	0.56 - 0.97
Item 3: Properties of Solutions				
Difficulty: Incorrect or DCI only vs DCI + SEP	0.96	0.40 - 1.53	1.95	1.05 - 2.85
Difficulty: DCI + SEP vs DCI + SEP + CC	0.21	-0.22 - 0.64	-0.35	-0.97 - 0.26
Discrimination	0.67	0.43 - 0.92	0.60	0.38 - 0.83
Item 4: States of Matter				
Difficulty: Incorrect vs DCI or CC	-6.66	-10.38 - -2.94	-11.30	-20.59 - -2.01
Difficulty: DCI or CC vs DCI + CC	-0.70	-1.30 - -0.11	-0.40	-1.08 - 0.28
Discrimination	0.39	0.16 - 0.63	0.28	0.04 - 0.51

Note. To determine the difficulty and discrimination parameters, each item was tested in its' own hybrid model using a 3PL model for multiple choice and PCM for the constructed response.

Bias

The AACR algorithms showed the possibility of bias for two of the three items.

Bias was not consistent across items.

So that bias is not inherent in the AACR algorithm.

Diversity in Research

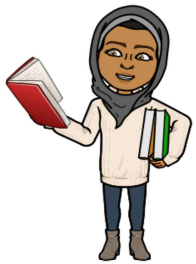
```
. tab HomeLanguage if inlist(LangCode,2,4,5)
```

HomeLanguage	Freq.	Percent	Cum.
A TERRIBLE SPANISH	1	0.48	0.48
ARABIC	2	0.96	1.44
ARABIC OR ENGLISH	1	0.48	1.91
BULGERIAN	1	0.48	2.39
CANTONESE	1	0.48	2.87
CANTONESE/ENGLISH	1	0.48	3.35
CHINESE	3	1.44	4.78
ENGLIS OR VIETNAMSE	1	0.48	5.26
ENGLISH AND A HELL OF A LOT OF GERMAN..	1	0.48	5.74
ENGLISH AND SPANISH I SPEAK BOTH QUIT..	1	0.48	6.22
ENGLISH, ASL	1	0.48	6.70
ENGLISH, I CAN UNDERSTAND SOME FILIPI..	1	0.48	7.18
ENGLISH, SOMETIMES GERMAN	1	0.48	7.66
ENGLISH/ TAGALOG	1	0.48	8.13
ENGLISH/FILIPINO	1	0.48	8.61
ENGLISH/GERMAN	1	0.48	9.09
ESPANOL	4	1.91	11.00
ESPAÑOL	2	0.96	11.96
FILIPINO	1	0.48	12.44
FILIPINO/ENGLISH	1	0.48	12.92
HUNGARIAN/GERMAN	1	0.48	13.40
I SPEAK ENGLISH. BUT I ALSO DO ALOT O..	1	0.48	13.88
KAPAMPANGAN (DIALECT IN FILIPINO)	1	0.48	14.35
MANDINGO	1	0.48	14.83
PUNJABI	1	0.48	15.31
SPANISH	170	81.34	96.65
TAGALOG	3	1.44	98.09
TAGALOG/ENGLISH	1	0.48	98.56
TAGALUG	1	0.48	99.04
VIETNAMESE	2	0.96	100.00
Total	209	100.00	

Other

RaceCode	Freq.	Percent	Cum.
White, non-hispanic	359	32.08	32.08
Hispanic	522	46.65	78.73
Black	113	10.10	88.83
Asian	65	5.81	94.64
Other	22	1.97	96.60
Multiracial	38	3.40	100.00
Total	1,119	100.00	

LangCode	Freq.	Percent	Cum.
English	690	72.40	72.40
Spanish	177	18.57	90.98
English & Spanish	54	5.67	96.64
Other	18	1.89	98.53
English & Other	14	1.47	100.00
Total	953	100.00	



Languages spoken in the homes of students who reported their race as one that was grouped into the category of “Asian”

HomeLanguage	Freq.	Percent	Cum.
-99	1	1.67	1.67
BANGALI	1	1.67	3.33
BANGLA	2	3.33	6.67
BENGALI	3	5.00	11.67
BULGERIAN	1	1.67	13.33
CANTONESE/ENGLISH	1	1.67	15.00
CHINESE	2	3.33	18.33
ENGLIS OR VIETNAMSE	1	1.67	20.00
ENGLISH	32	53.33	73.33
ENGLISH/ TAGALOG	1	1.67	75.00
ENGLISH/FILIPINO	1	1.67	76.67
FILIPINO	1	1.67	78.33
KAPAMPANGAN (DIALECT IN FILIPINO)	1	1.67	80.00
KOREAN	1	1.67	81.67
NEPALI	1	1.67	83.33
PUNJABI	1	1.67	85.00
TAGALOG	3	5.00	90.00
TAGALOG/ENGLISH	1	1.67	91.67
THAI	1	1.67	93.33
THAI LAUGUAGE	1	1.67	95.00
URDU	1	1.67	96.67
VIETNAMESE	2	3.33	100.00
Total	60	100.00	

Analysis

We created a binary variable indicating whether the machine score was

- Lower than the human score.
- Higher than the human score.

The odds of student *i* being scored lower (or higher) by the machine than the human raters:

$$\frac{p}{1-p} = \exp(\beta_0 + \beta_1 Predictor_i + \beta_2 MultipleChoice_i)$$

And the probability score from AACR was that the score of student *i* was correct was regressed similarly:

$$P(Correct_i) = \beta_0 + \beta_1 Predictor_i + \beta_2 MultipleChoice_i + \varepsilon_i$$

Analysis (Multivariate)

The odds of student *i* being scored lower (or higher) by the machine than the human raters:

$$\frac{p}{1-p} = \exp(\beta_0 + \beta_1 RE(Hispanic)_i + \beta_2 RE(Black)_i + \beta_3 RE(Asian)_i + \beta_4 RE(Other)_i + \beta_5 RE(Multiple)_i + \beta_6 Lang(Spanish)_i + \beta_7 Lang(Eng\&Spanish)_i + \beta_8 Lang(Other)_i + \beta_9 Lang(Eng\&Other)_i + \beta_{10} MultipleChoice_i + \beta_{11}(NumWords)_i)$$

And the probability score from AACR was that the score of student *i* was correct was regressed similarly:

$$P(Correct_i) = \beta_0 + \beta_1 RE(Hispanic)_i + \beta_2 RE(Black)_i + \beta_3 RE(Asian)_i + \beta_4 RE(Other)_i + \beta_5 RE(Multiple)_i + \beta_6 Lang(Spanish)_i + \beta_7 Lang(Eng\&Spanish)_i + \beta_8 Lang(Other)_i + \beta_9 Lang(Eng\&Other)_i + \beta_{10} MultipleChoice_i + \beta_{11}(NumWords)_i + \varepsilon_i$$

When the Machine Scored Lower

Log Odds from Models When the Machine Scored Lower

	Univariate						Multivariate					
	Item 2:		Item 3:		Item 4:		Item 2:		Item 3:		Item 4:	
	Odds Ratio	S.E.	Odds Ratio	S.E.	Odds Ratio	S.E.	Odds Ratio	S.E.	Odds Ratio	S.E.	Odds Ratio	S.E.
Female	0.49 *	0.14	0.85	0.25	0.86	0.29	0.51 *	0.17	0.95	0.32	0.98	0.38
Race (White non-Hispanic as Reference)												
Hispanic	0.83	0.25	1.11	0.42	1.10	0.43	0.72	0.31	1.37	0.66	0.41	0.25
Black	0.34	0.26	0.88	0.54	1.67	0.95	0.21	0.22	0.65	0.46	1.70	1.00
Asian	1.59	0.80	4.97 **	2.59	0.51	0.54	0.90	0.65	6.73 **	4.32	0.82	0.91
Other	3.36	2.10	empty		2.04	2.29	3.87 *	2.55	Empty		2.05	2.40
Multiple Races	1.35	0.89	1.64	1.39	2.00	1.64	1.43	1.00	1.52	1.30	2.19	1.80
Home Language (English Only as Reference)												
Spanish	0.84	0.34	0.79	0.35	1.91	0.76	0.92	0.49	0.66	0.35	4.73 *	2.96
English and Spanish	1.08	0.68	0.54	0.41	1.15	0.89	1.19	0.85	0.47	0.38	2.73	2.43
Other	0.76	0.80	2.70	2.39	Empty		0.56	0.61	0.77	0.82	Empty	
English and Other	Empty		Empty		Empty		Empty		Empty		Empty	
Number of Words	1.36 **	0.15	1.16	0.14	1.19	0.15	1.48 **	0.19	1.08	0.18	1.31	0.20
Standardized MC	1.44 *	0.26	1.36	0.21	1.02	0.20	1.01	0.20	1.09	0.24	1.14	0.28

Note. A binary variable indicating whether the machine scored lower than humans was created. Due to potential correlations between race and language, each was included as a coviariate in a logistic regression controlling for students' standardized scores on the pretest portion to provide the odds ratios seen in the table. *p < .05, **p < .01, **p <.001

When the Machine Scored Higher

Log Odds from Models When the Machine Scored Higher

	Univariate						Multivariate					
	Item 2:		Item 3:		Item 4:		Item 2:		Item 3:		Item 4:	
	Odds Ratio	S.E.	Odds Ratio	S.E.	Odds Ratio	S.E.	Odds Ratio	S.E.	Odds Ratio	S.E.	Odds Ratio	S.E.
Female	1.78	0.67	1.14	0.48	0.82	0.27	1.63	0.79	1.07	0.49	1.04	0.39
Race (White non-Hispanic as Reference)												
Hispanic	1.00	0.43	1.38	0.70	0.97	0.41	1.02	0.60	0.88	0.63	1.47	0.80
Black	1.16	0.74	1.05	0.87	1.88	1.06	1.43	1.06	0.93	0.78	1.49	0.96
Asian	Empty		1.59	1.31	6.44 **	3.52	Empty		2.30	1.93	6.74 **	4.38
Other	1.44	1.54	3.59	4.08	4.48	4.01	2.46	2.89	3.63	4.87	4.99	4.75
Multiple Races	Empty		1.72	1.98	0.99	1.02	Empty		1.65	1.85	0.97	1.08
Home Language (English Only as Reference)												
Spanish	0.58	0.38	1.06	0.63	0.81	0.40	0.45	0.34	1.02	0.82	0.87	0.52
English and Spanish	0.62	0.64	0.63	0.67	1.03	0.78	0.22	0.33	0.69	0.84	1.12	0.91
Other	Empty		Empty		14.16 **	13.09	Empty		Empty		4.80	4.58
English and Other	Empty		Empty		2.46	3.50	Empty		Empty		0.91	1.39
Number of Words	2.23 ***	0.28	1.74 ***	0.28	0.75	0.12	2.61 ***	0.45	1.94 ***	0.37	0.81	0.16
Standardized MC	0.84	0.13	1.72 *	0.37	0.75	0.14	0.56 **	0.12	1.17	0.30	0.57 *	0.13

Note. A binary variable indicating whether the machine scored higher than humans was created. Models were tested with each covariate tested individually and together. *p < .05, **p < .01, **p <.001

Certainty of Prediction

Predicted Probability of a Correct Response by Demographic Data

	Univariate						Multivariate					
	Item 2:		Item 3:		Item 4:		Item 2:		Item 3:		Item 4:	
	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.
Female	0.00	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.01
Race (White non-Hispanic as Reference)												
Hispanic	-0.01	0.01	0.00	0.01	0.00	0.01	-0.01	0.01	0.00	0.02	0.00	0.01
Black	-0.01	0.01	0.01	0.02	0.00	0.02	-0.01	0.01	0.03	0.02	0.00	0.02
Asian	-0.02	0.02	-0.06 *	0.02	-0.05 *	0.02	0.01	0.02	-0.07 *	0.03	-0.07 **	0.03
Other	-0.02	0.02	-0.01	0.04	-0.05	0.04	-0.02	0.02	0.00	0.04	-0.04	0.03
Multiple Races	-0.03	0.02	0.04	0.03	0.00	0.03	-0.01	0.02	0.05	0.03	0.00	0.02
Home Language (English Only as Reference)												
Spanish	0.00	0.01	0.02	0.01	-0.01	0.01	0.00	0.01	0.02	0.02	-0.01	0.02
English and Spanish	-0.01	0.02	-0.01	0.02	-0.01	0.02	0.00	0.02	-0.01	0.03	-0.01	0.02
Other	0.00	0.02	-0.02	0.03	-0.03	0.04	-0.01	0.02	0.03	0.03	0.03	0.04
English and Other	0.02	0.02	0.03	0.06	0.00	0.04	0.01	0.02	0.11	0.07	0.05	0.05
Number of Words	-0.04 ***	0.00	-0.02 ***	0.00	0.03 ***	0.01	-0.05 ***	0.00	-0.02 *	0.01	0.03 ***	0.01
Standardized MC	-0.01 ***	0.00	-0.01	0.01	0.02	0.01	0.00	0.00	0.00	0.01	0.02 **	0.01

Note. A binary variable indicating whether the machine scored higher than humans was created. Models were tested with each covariate tested individually and together. *p < .05, **p < .01, **p <.001

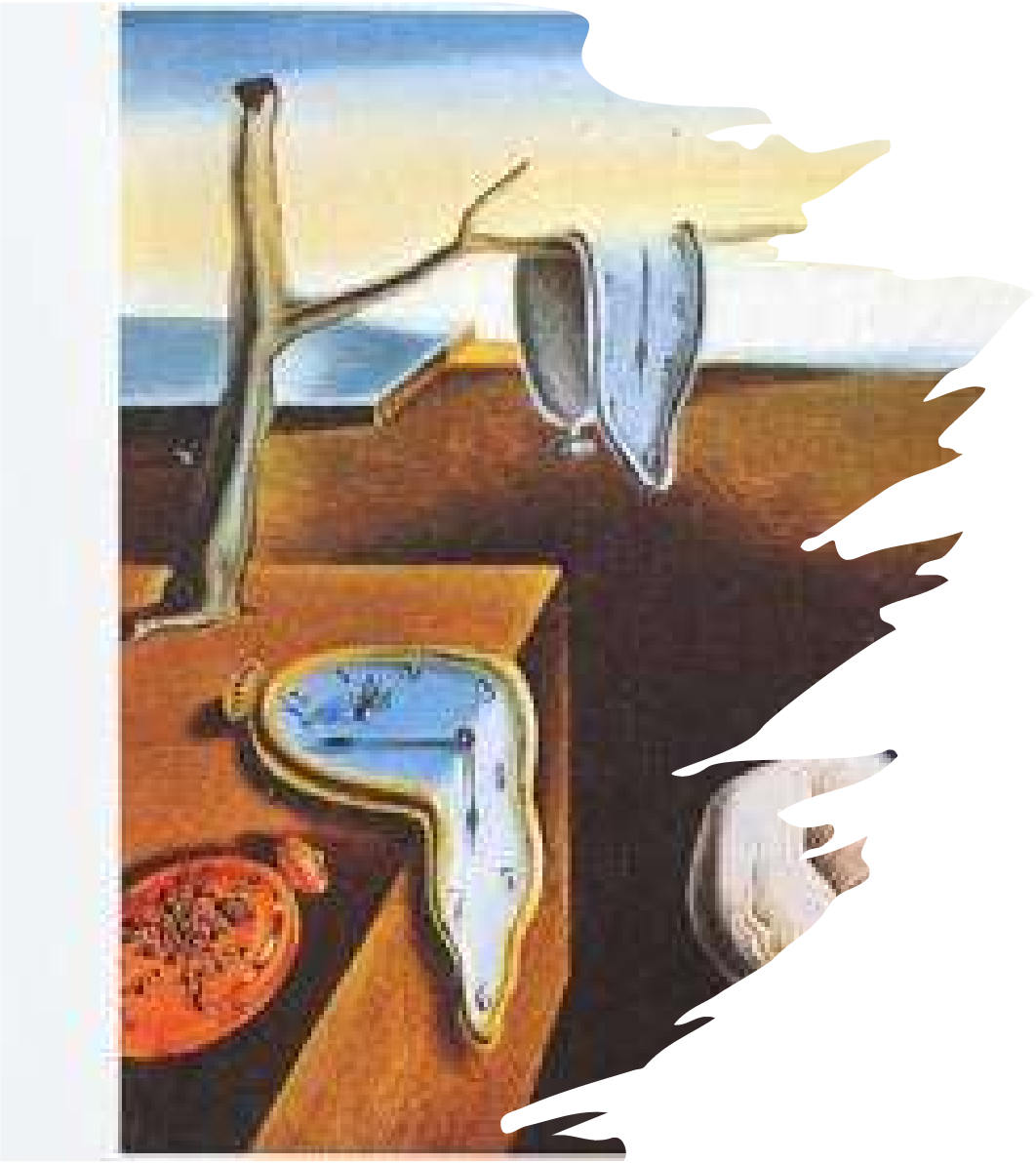
Machine and Human Disagreed (High and Low Combined)

Log Odds from Models When the Machine and Humans Disagreed

	Univariate						Multivariate					
	Item 2:		Item 3:		Item 4:		Item 2:		Item 3:		Item 4:	
	Odds Ratio	S.E.	Odds Ratio	S.E.	Odds Ratio	S.E.	Odds Ratio	S.E.	Odds Ratio	S.E.	Odds Ratio	S.E.
Female	0.78	0.17	0.92	0.24	0.83	0.20	0.70	0.19	0.98	0.29	0.97	0.27
Race (White non-Hispanic as Reference)												
Hispanic	0.89	0.23	1.22	0.39	1.02	0.31	0.74	0.27	1.21	0.51	0.78	0.33
Black	0.63	0.30	0.93	0.48	1.88	0.80	0.53	0.31	0.73	0.41	1.66	0.75
Asian	1.02	0.49	4.55 **	2.25	3.26 *	1.54	0.59	0.43	7.97 **	5.19	3.47 *	1.89
Other	2.93	1.69	0.98	1.07	3.82	2.82	3.87 *	2.47	0.90	1.07	4.05	2.97
Multiple Races	0.85	0.55	1.78	1.34	1.54	1.01	0.84	0.59	1.61	1.22	1.53	1.05
Home Language (English Only as Reference)												
Spanish	0.73	0.26	0.86	0.32	1.32	0.43	0.73	0.34	0.73	0.34	2.10	0.92
English and Spanish	0.91	0.50	0.54	0.35	1.07	0.62	0.77	0.54	0.50	0.34	1.68	1.09
Other	0.46	0.47	1.64	1.45	5.78	5.57	0.50	0.55	0.31	0.34	2.59	2.57
English and Other	Empty		Empty		1.29	1.55	Empty		Empty		0.73	0.90
Number of Words	1.79 ***	0.18	1.43 **	0.17	1.00	0.13	1.95 ***	0.23	1.43 *	0.21	1.11	0.18
Standardized MC	1.18	0.15	1.53 **	0.21	0.86	0.12	0.81	0.13	1.12	1.12	0.78	0.13

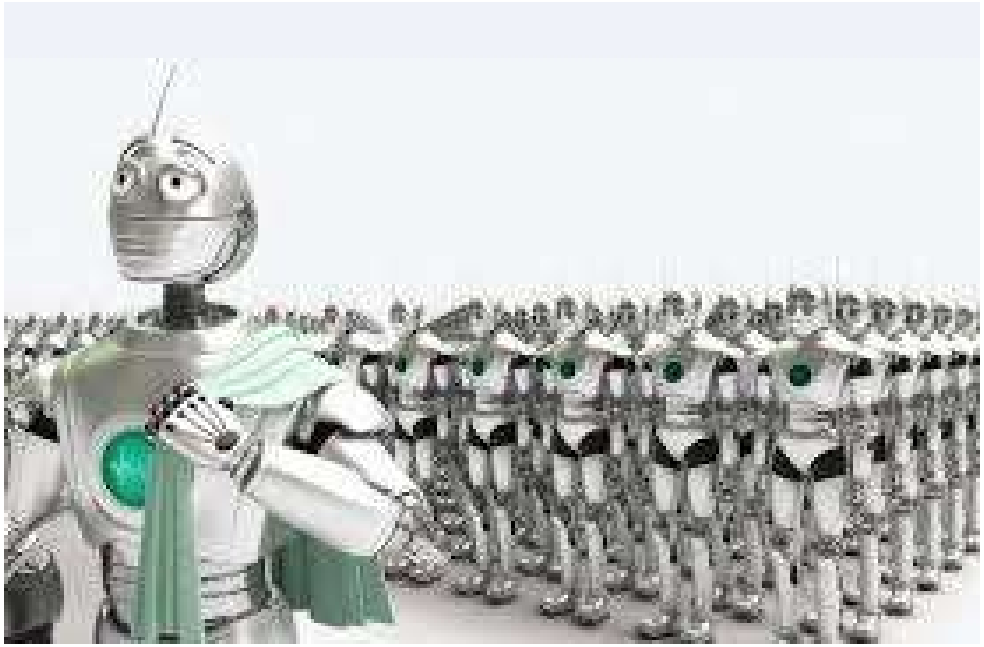
Note. A binary variable indicating whether the machine scored lower than humans was created. Models were tested with each covariate tested individually and together. *p < .05, **p < .01, **p <.001

Discussion & Conclusion



Maintaining Human-to-Human Agreement

- Human scoring
  - Extensive training exercises
  - Scheduling inconsistencies
  - Training new raters
  - Fluctuations in inter-rater agreement

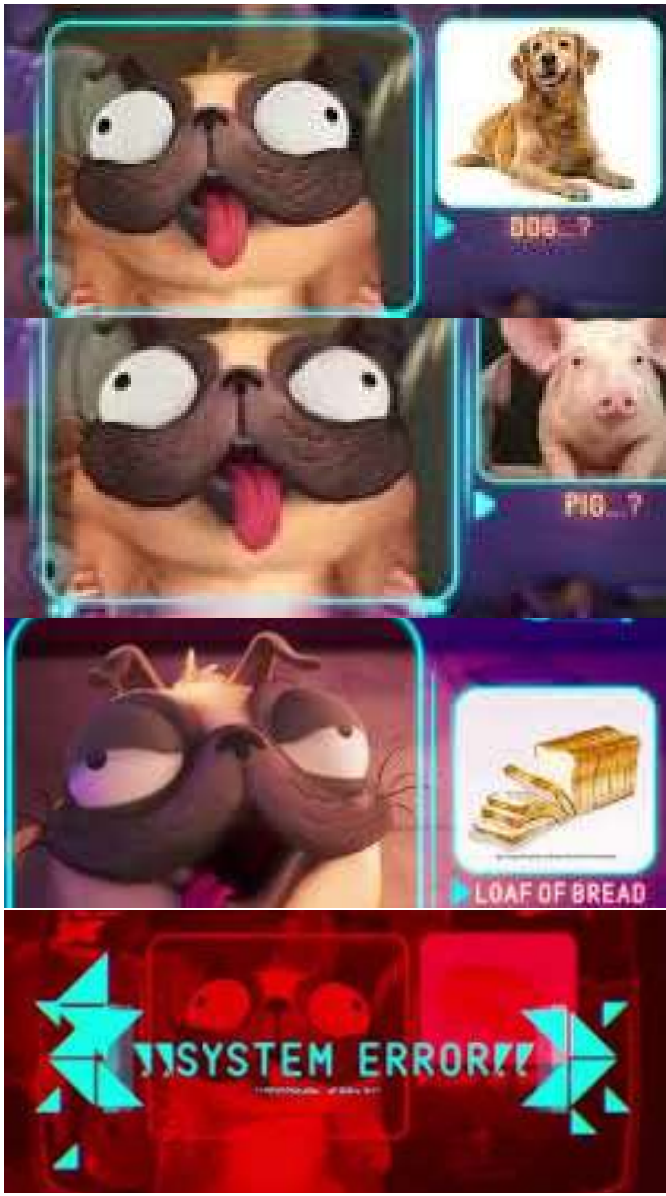


Efficiency

- Training humans was time and labor intensive.
- Once completed, the algorithms scored remaining responses much faster than human raters were able.

Setbacks

- Although the bias was not consistent across items, the machine algorithms showed bias on some items.
- Changes to difficulty and discrimination parameters may require adjustments in assessment development.
- Items where scoring categories are underrepresented are more difficult to score.



Conclusion

- Automated analysis is an effective method for researchers to reduce time and labor costs in scoring student constructed responses and should be used with appropriate caution.
  - Obtain high agreement between humans.
  - Consider the importance of difficulty and discrimination parameters in the study.
  - Explore each algorithm for bias.
- At the classroom level, teachers would need to search existing item banks for items already being used with machine scoring as large training sets are needed.

References

Balfour, S. P. (2013). Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review. Research and Practice in Assessment, 8, 40-48.

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. Educational Measurement: Issues and Practice, 38(2), 53-67. https://doi.org/10.1111/emip.12253.

Lee, N. T. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*.

Liu, O. L., Brew, C., Blackmore, J., & Gerard, L. (2014). Automated scoring of constructed response science items: Prospects and obstacles. Educational Measurement-Issues and Practices, 33(2), 19–28. https://doi.org/10.1111/emip.12028.

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H.-S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. Educational Assessment, 23(2), 121-138.

National Research Council. (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. National Academies Press.

National Research Council. (2014). Developing assessments for the next generation science standards. National Academies Press.

References

Nehm, R. H., & Haertig, H. (2012). Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software. Journal of Science Education and Technology, 21(1), 56-73.

NGSS Lead States. (2013). Next Generation Science Standards: For States, By States. Washington, DC: The National Academies Press.

Raykov, T., & Marcoulides, G. A. (2018). A course in item response theory and modeling with Stata. College Station, TX: Stata Press.

Schneider, B., Krajcik, J., Lavonen, J., Salmela-Aro, K., Klager, C., Bradford, L., ... & Bartz, K. (2022). Improving science achievement—Is it possible? Evaluating the efficacy of a high school chemistry and physics project-based learning intervention. Educational Researcher, 51(2), 109-121.

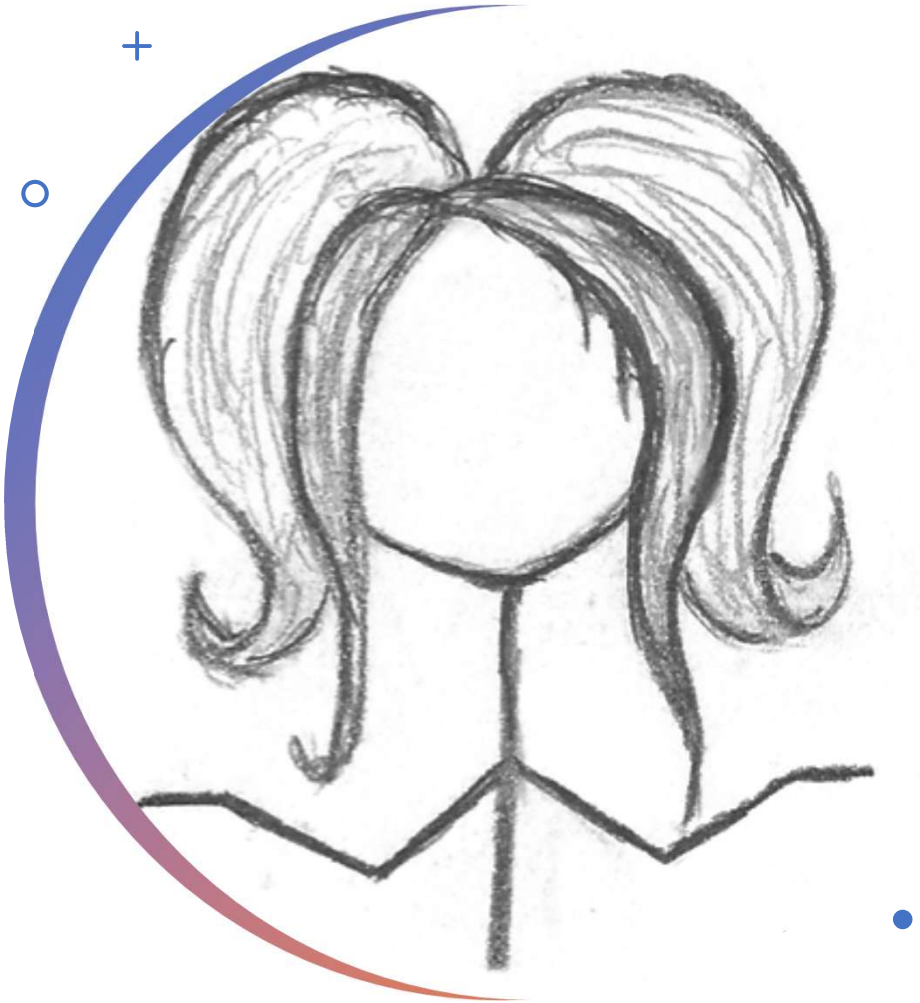
Yapo, A., & Weiss, J. (2018, January). Ethical implications of bias in machine learning. In Proceedings of the 51st Hawaii International Conference on System Sciences.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020c). Applying machine learning in science assessment: a systematic review. Studies in Science Education, 56(1), 111-151.

OISE: 1545684

Using Machine Learning to Score Multi-Dimensional Science Assessments

- Sarah Maestrales
- Xiaoming Zhai
- Israel Touitou
- Quinton Baker
- Joseph Krajcik
- Barbara Schneider



Thanks so much for having me!

Sarah Maestrales
maestral@msu.edu