How useful can a machine be in providing reliable and valid diagnostic information about writing to teachers and learners?

Automated essay scoring (AES)

Long history (initiator Page, 1966)

- linguistic parsing (parts of speech, word identification, sentence structures) analysed and correlated to predict a score
- success or otherwise was judged on how well this score matched human markers' scores

Developments over the decade draw on computational power of computers

- Method of judging success has not changed

All systems use a 'training set' of essays written to one topic that have been marked by humans to develop a scoring model that is used subsequently to score new, unknown essays written to the same topic.

Automated essay scoring (AES)

- Many claims of the benefits of AES
- Increasing evidence of the consistency of the scoring models (i.e. claims that machine scoring can replicate human markers scoring at least as well as human markers can replicate each other's scores)

However, very controversial within research community:

- writing is a communicative act between humans
- computers do not think and cannot analyse quality of thought the way a human can

Automated essay scoring (AES)

- Important to examine and report the quality of score replication for all new contexts in which AES is proposed.

- AES is not a generic, one size fits all writing scoring machine. AES utilises task – specific scoring models, unique to each assessment context.
  - For this context, two scoring models were built.

Research questions

Research questions

- How effective is a machine at replicating the scores of human raters across different writing tasks?
- Are there differences in its effectiveness across different task types?
- Are there differences in its effectiveness across different marking criteria? Do these differ across task types?
- What features characterise writing that the machine was unable to score or had scores that were discrepant from human scores?

Core skills Profile for Adults

http://www.acer.edu.au/cspa/online-writing-assessment

The assessment context

Collecting evidence about student and group learning to

- estimate current level of attainment
- improve learning

Good assessment practices

- fair, appropriate, relevant, valid
  - accessible tasks suited to target test takers [ACSF Levels 2-4]
  - authentic, every day and functional tasks
  - curriculum aligned assessment criteria

Provide useful reporting information that would be of value to teachers and learners. Results that are

- accurate, valid, and reliable
- draw on a range of evidence
- give an objective view
- manageable and timely

Assessment development process

- A series of assessment tasks developed
- Tasks piloted with groups of students from a variety of organisations
- Most viable tasks selected to go forward to a full scale trial
- Marking criteria developed from pilot scripts and with reference to ACSF indicators and focus areas
- Tasks and marking guide trialed
- Scripts double-marked and adjudicated by experienced markers
- Development of measurement and reporting scale
- Development of the automated marking system (training the software)
- Development of test delivery and reporting platform

The assessment tasks

- Task 1: requires the writer to explain a problem and offer a possible solution in a note to a known audience in order to induce actions in that audience.

- Task 2: requires the writer to make a request, related to the context of the first task, to an unknown official audience

Types of responses to Task 1

Dear (neighbours name)

I am very sorry to have to approach this matter and i thought maybe a letter might help.

I am having hard time managing to get to sleep, am being woken and also am finding it very hard to concentrate when studing while your dog, which i do like, is barking. He barks constanly. Maybe i can help you find something that might mak him a little calmer, as im sure it must irritate you also.

I hope i am not overstepping our boundries and that we can come to some sort of solution for this problem.

A few ideads:

a chew toy

keep him inside a night

more regular walks

another dog

bones

Thank you for taking the time to hear me out.

Sincerly

Your neighbour.

Types of responses

Dear neighbour,

Your poor old dog is missing you when you are at work. She is barking constantly. I figure she is bored and lonely.

Perhaps we could come to some solution to help her feel less bored and alone. I have some ideas, come over for a cuppa and we'll work something out.

See you soon,

Your neighbour.

Types of responses

To whom it may concern,

I am writing to you about your dog's constant barking at night. Would you kindly take your dog inside during this time as i wake up early in the morning for work.

Your sincerly,

Your neighbour

Types of responses

Stick your dog up ya clacka

Developing the marking criteria

- Analytic, criterion-referenced guide
- Eight marking criteria defined by an assessment focus and elaborated by between two and four scoring categories
- identification of assessment criteria through curriculum documents
- a selection of scripts were used to develop and describe ordered scoring categories to ensure the marking guide reflects the demands of the task
- Scripts that exemplify the categories of each criterion
- The criteria and a selection of the pilot scripts ('exemplars') constitute the method for obtaining consistent, valid and reliable judgements when marking student writing.

The marking criteria

Marking Criteria	Assessment focus	Score range
Audience and Purpose	The writer's capacity to orient and engage the reader by providing relevant information and using a consistent and appropriate register	0–3
Spelling	The writer's capacity to generate and spell correctly words with a range of difficulties	0–3
Language choices	The range and precision of language	0–2
Sentence structure	The production of grammatically correct, structurally sound and meaningful sentences	0–2
Punctuation of Sentences	The correct use of appropriate punctuation of sentences (capital letters to start sentences and full stops to finish)	0–1
Punctuation within Sentences	The correct use of appropriate punctuation within sentences	0–2
Text cohesion	The production of cohesive text, rendered navigable by text connectives	0–2
Quality of ideas	The generation of relevant, extended and elaborated ideas	0–3

The marking guide

1. Purpose and Audience

Skill focus:	The writer's capacity to orient and engage the reader by providing relevant information and by consistently using an appropriate register.

2. Spelling

Skill focus:	The writer's capacity to generate and spell correctly words with a range of difficulties.

3. Language choices

Skill focus:	The range and precision of language choices.

4. Punctuation of sentences

Skill focus:	The correct use of appropriate punctuation of sentences.

Category score	Category descriptor	Additional information	Exemplars
0	– there is little correct use of capital letters to start sentences OR full stops to end sentences (or attempted sentences) *sentence punctuation is minimal and of little assistance to the reader*	– sentence punctuation includes: – capital letters to begin sentences – full stops, question marks and exclamation marks to end sentences – sentence punctuation may be appropriate to form, e.g. capital letters after colon in Q&A series	**Task A:** All day Dog training **Task B:** Hi your dog Legal action
1	– sentences are accurately punctuated (must have at least 2 sentences) – one or two errors are allowed in a longer text *sentence punctuation assists reading*		**Task A:** Quite disruptive Sleep in **Task B:** Constantly barking Peace and quiet

5. Punctuation within sentences

6. Sentence structure

7. Text cohesion

Skill focus:	The production of cohesive text, rendered enviable by text connectives.

8. Quality of ideas

Skill focus:	The generation of relevant, extended and elaborated ideas.
Notes:	– When a sentence is extended, more information is provided about the sentence's basic idea. – When a sentence is elaborated, the basic idea and/or its extensions are embellished with further explanation or detail such as background information, an example or a consequence or result.

Category score	Category descriptor	Additional information	Exemplars
0	– has few plausible ideas – does not use ideas offered in the instructions or uses inappropriate ideas	– may copy ideas from instructions	
1	– uses ideas offered in the instructions OR – uses 1 or 2 of own ideas – gives little extension or elaboration of ideas	– uses minimal and simplistic ideas	**Task A:** All day Dog training **Task B:** Hi your dog Legal action Constantly barking
2	– uses own ideas AND/OR – uses ideas offered in the instructions – gives some relevant extension or elaboration of ideas		**Task A:** Quite disruptive
3	– uses own ideas AND/OR – uses ideas offered in the instructions – effectively extends, or elaborates on, ideas		**Task A:** Sleep in **Task B:** Peace and quiet

Links to the ACSF

No.	Criteria	ACSF indicators and performance foci
1	**Purpose and audience**	05: Range 05: Audience and purpose 05: Register 06: Grammar
2	**Quality of ideas**	05: Range 05: Structure & cohesion
3	**Text cohesion**	05: Structure & cohesion 06: Grammar
4	**Language choices**	06: Vocabulary 06: Grammar
5	**Sentence structure**	06: Grammar
6	**Punctuation OF sentences**	06: Punctuation
7	**Punctuation WITHIN sentences**	06: Punctuation
8	**Spelling**	06: Spelling

Marking operations

- Two separate operations
- Double blind marking
- Third marker to adjudicate discrepant human scores
- Marking on-screen

Score array

Student id	Answer	Audience	Quality of Ideas	Text Cohesion	Language Choices	Sentence structure	Punc of sentences	Punc in sentences	Spelling	TOTAL
	max score	**3**	**3**	**2**	**2**	**2**	**1**	**2**	**3**	**18**
1	hello neighbour your dog is keeping me up with its barking can you pleaSE PUT IT IN SIDE SO IT WILL STOP BARKING THANKYOU.	1	2	1	0	1	0	0	1	**6**
3	Dear Neighbour, It has come to my attention that at late at night and early hours of the morning, your dog seems to constantly bark. Im sorry to tell you but it does get really annoying, and i was wondering if somehow you could get it to slow down or stop, i know it will be hard but its a suggestion that i would like if you took into consideration. Thanks heaps.	2	2	2	1	2	1	1	3	**14**
5	dear patrents of this naghberhood it has come to my atention that your beloved canine has been a nusence to this naughb or has it keeps me up at all hours of the night i wish to ask for a selothin that we can both agree on to resolve the promblem, kind regards.	1	2	1	2	0	0	0	2	**8**
9	To whom it may concern, It has come to my attention that your pet dog seems to constantly bark throughout day and night. It would not be a problem but since there are young children that live next door it becomes disruptive and is hard for the children to get a good nights rest. If possible would your dog be able to be brought in during night and kept in the laundry if unwanted within the house. This would be deeply appriciated. Thnak-you for your time and consideration. Your friendly neighbour	1	3	2	2	2	1	1	3	**15**
11	Your dog is very annoying and i would like it if You Keep it inside	1	1	1	0	0	0	0	1	**4**
12	Talk to them nicely about the dog being a nucense. If not complain to the council.	0	1	1	1	0	0	0	1	**4**

Rasch analysis

- Check function of marking guide on each task
  - Expected score curves for each criterion
  - Item characteristic score curves for each criterion and category
- Test characteristics: by each task and for combined tasks

Measurement and reporting

- Performance scale
- Performance profile
- ACSF levels mapped to profile and scale
- Response pattern for both tasks

Reporting – need to emphasise that this is the purpose

Reporting: the process of communicating information about student or group learning, including

- level of attainment
- progress made

Reports should

- meet the institution's reporting requirements
- be easy to understand
- be time efficient and easy to access
- show progress against learning objectives
  - what has been taught
  - what the student has learnt
  - what the student doesn't know or can't do yet.
- show what is expected at particular levels or stages or course outcomes
- have provision for monitoring over time: individual, class, organisation
- have supporting evidence for judgements about attainment and progress
- be constructive
- be a basis for discussion between teacher and student

Electronic reporting

Jayden.pdf

Jayden response.pdf

Test GRT.pdf

Test GRG.pdf

'Training' the automated scoring system

- Include at least 300 training scripts
- Provide sufficient coverage across each category score point, including the tails
- Include multiple markers
- Ensure the markers are well calibrated and scores are consistent
- Include an additional set of 50 "validation" scripts for which the scores are withheld from the training
- Use typed scripts

How effective is a machine at replicating the scores of human raters across different writing tasks?

- Total scores for each task
- Correlations , Mean and SD, agreement rates
  - Marker 1 vs Marker 2
  - Marker 1 vs machine
  - Marker 2 vs machine

Total score correlations

	Task 1 Note	Task 2 Letter
	N = 334	N = 332
Marker 1 and 2	0.88	0.85
Marker 1 and Machine	0.89	0.86
Marker 2 and Machine	0.89	0.88

Mean and SD

	Task 1 Note		Task 2 Letter	
	Mean	SD	Mean	SD
Marker 1	10.59	4.43	10.74	4.37
Marker 2	10.71	4.74	11.03	4.56
Machine	11.39	4.65	11.87	4.41

Total score agreement rates

		Marker 1 and Marker 2	Marker 1 and machine	Marker 2 and machine
Exact	Task 1	20%	21%	16%
Exact	Task 2	20%	16%	23%
adjacent	Task 1	33%	31%	33%
adjacent	Task 2	29%	29%	32%
discrepant	Task 1	22%	21%	26%
discrepant	Task 2	21%	20%	19%
discrepant >2	Task 1	25%	26%	24%
discrepant >2	Task 2	30%	35%	26%

Are there differences in the machine's effectiveness across different marking criteria? Do these differ across task types?

- Criteria scores for each task

- Correlations and agreement rates
  - Marker 1 and Marker 2
  - Marker 1 and machine
  - Marker 2 and machine

Task 1 correlations by criteria

Task 1 No. = 334	Audience	Quality of Ideas	Text Cohesion	Language Choices	Sentence Structure	Punc OF Sentences	Punc IN Sentences	Spelling
Marker 1 and 2	0.64	0.78	0.63	0.65	0.66	0.71	0.65	0.70
Marker 1 and machine	0.64	0.78	0.68	0.74	0.65	0.72	0.66	0.73
Marker 2 and machine	0.69	0.78	0.65	0.74	0.66	0.73	0.71	0.73

Task 2 correlations by criteria

Task 2 No. = 322	Audience	Quality of Ideas	Text Cohesion	Language Choices	Sentence Structure	Punc Of Sentences	Punc In Sentences	Spelling
Marker 1 and 2	0.67	0.62	0.50	0.63	0.63	0.71	0.58	0.59
Marker 1 and Machine	0.69	0.70	0.58	0.67	0.60	0.69	0.61	0.66
Marker 2 and Machine	0.72	0.78	0.66	0.68	0.57	0.77	0.70	0.70

Task 1 Criteria Agreement

Task 1 N=334		Audience	Quality of Ideas	Text Cohesion	Language Choices	Sentence Structure	Punc Of Sentences	Punc In Sentences	Spelling
Exact	Markers 1 and 2	60%	69%	67%	63%	65%	86%	68%	63%
	Marker 1 and Machine	64%	72%	75%	67%	68%	87%	72%	70%
	Marker 2 and Machine	64%	69%	70%	71%	66%	87%	74%	67%
Adjacent	Markers 1 and 2	38%	31%	32%	36%	35%	14%	31%	37%
	Marker 1 and Machine	35%	28%	24%	33%	32%	13%	28%	30%
	Marker 2 and Machine	36%	31%	29%	29%	33%	13%	25%	32%
Discrepant	Markers 1 and 2	2%	0%	1%	1%	0%	0%	1%	1%
	Marker 1 and Machine	1%	0%	1%	0%	0%	0%	1%	0%
	Marker 2 and Machine	0%	0%	1%	0%	1%	0%	0%	1%

Task 2 Criteria Agreement

Task 2 N=322		Audience	Quality of Ideas	Text Cohesion	Language Choices	Sentence Structure	Punc Of Sentences	Punc In Sentences	Spelling
Exact	Markers 1 and 2	58%	57%	57%	63%	66%	87%	62%	61%
	Marker 1 and Machine	59%	62%	63%	69%	67%	86%	66%	66%
	Marker 2 and Machine	68%	71%	76%	70%	65%	90%	70%	65%
Adjacent	Markers 1 and 2	41%	40%	42%	37%	34%	13%	35%	36%
	Marker 1 and Machine	39%	36%	37%	31%	32%	14%	33%	34%
	Marker 2 and Machine	31%	29%	22%	30%	35%	10%	30%	34%
Discrepant	Marker s 1 and 2	1%	3%	1%	0%	0%	0%	2%	3%
	Marker 1 and Machine	2%	1%	1%	0%	1%	0%	1%	0%
	Marker 2 and Machine	1%	0%	1%	0%	1%	0%	0%	1%

Writing where machine scoring was discrepant from human scoring

Dear neighbour,

This letter is in regards to your dogs constant barking, i am your neighbour that lives on your right. Your dog has seemed to become a nuisance because of its constnat barking.

I would very much apreciate it if you would try to reduce your dogs amount of barking during the day and night as it has become a problem. The barking is interfering with the enjoyment of my property and is something that must be stopped before i take this further.

I have come up with a soloution to fix this problem, i strongly suggest that you put the dog inside of your house or in the garage to minimize the noise.

Kind regards, your neighbour.

Non-scoreable scripts

- Off-topic
- Too short or insufficiently developed
- Major syntax errors
- Copy of the prompt
- Repetitious
- Too many unknown words

Some non-scoreable scripts

- hi, my name is jaycob grieve i live next door and i am justt letting you know about your dog its conatantly barking i have work realy early in the morning if there is anything you could do can yu please it will be very appreciated if possible can you please take your dog inside or find somehting to fix the problem i am an early starter and im always tired in the morning it will be appreciated for anything to help your dog ill be free to help to resolve the problem thankyou jaycob.

- you nkaw ferst iwell caolto the police my neighbour dog's befor hert the pepole do any solution

- Stick your dog up ya clacka

Conclusions – total scores

- Little difference in the quality of machine scoring between tasks
- Correlations are high on both tasks
- Machine produces a marginally higher mean score on both tasks (less than 1 on an 18pt scale)
- Machine can replicate the scores of one of the human markers at least as well as the markers can replicate each other's scores
- Human markers show slightly more discrepancy with each other than the machine and human markers shows
- Review process necessary (as with all assessments and exams)

Conclusions - criteria

❖On all but one criteria across both tasks, machine scoring can replicate the human markers at least as well as human markers can replicate each other's scores.

❖On both tasks, agreement rates across all criteria between the machine and human markers are at least as high as those between the human markers (between 96% - 100% exact and adjacent scoring.

❖On both tasks, machine scoring does not seem to be privileging the mechanics of writing elements over the context and ideas (cognition) elements.

❖ *The quality of replication by human markers is less on Task 2 (formal writing) for Audience and Purpose, Quality of Ideas and Text Cohesion criteria than it is for Task 1 (informal note)*

Summary of features

- Delivered online. No software needed.
- Valid
- Provides a range of useful information
  - diagnostic, summative and formative
  - student, group and organisation level
- Reports are available immediately: viewed, printed and/or downloaded
- Provides reliable and consistent marking over time
- Relies on high quality human marking to develop the scoring models
- Availability of response allows easy review of scores
- Does not replace a teacher or instructor