ABC NEWS

FACT CHECK

Fact file: Can a computer mark NAPLAN essays better than a teacher?

Posted Wed 15 Nov 2017 at 6:21am, updated Mon 16 Jul 2018 at 4:59pm

Some call it "robo-marking" and it's about to be put to the test. Next year, for the first time, essays written by school children as part of their national literacy and numeracy tests will be marked by a computer.

Teachers will also mark them as a back up, but if trials go according to plan, fully automated testing could be a reality within a few years, when all essays for the National Assessment Program — Literacy and Numeracy (NAPLAN) would be marked primarily by a computer.

The move to automated essay scoring will allow scores to be returned within three weeks rather than the current three months, helping teachers respond to students' learning needs sooner, according to the Australian Curriculum Assessment and Reporting Authority (ACARA), the body that oversees the tests.

Computers may well be suited to marking multiple-choice questions, but the prospect of machines marking essays, including stories, has attracted criticism, with some teachers and technology experts arguing computers cannot mark essays in the sentient way that humans can.

ACARA says its research, trials and analysis show computers are just as good as humans, if not better, at marking essays.

How does a computer mark an essay? And what did ACARA's research involve? RMIT ABC Fact Check explains.

NAPLAN: The low-down

NAPLAN is an annual assessment of students in Years 3, 5, 7 and 9 to test their skills in reading, writing, spelling, grammar, punctuation and numeracy.

Scores indicate whether a student is performing above, at or below the national minimum standard. They allow schools and governments to track overall student progress.

For the literacy test, students write an essay in styles known as persuasive (opinion writing) or narrative (storytelling).

Each state and territory is responsible for scoring the tests.

Teachers mark them in accordance with guidelines that take into account 10 aspects: audience, text structure, ideas, character and setting (narrative) and persuasive devices (persuasive), vocabulary, cohesion, paragraphing, sentence structure, punctuation and spelling.

How does automated essay scoring work?

Automated essay scoring (AES) is a computer system that uses algorithms designed to emulate human marking.

Before the system marks essays it must be 'trained' to score characteristics of writing, such as fluency, grammar and construction.

Also, using natural language processing (a component of artificial intelligence), the system can be programmed to score linguistic features, such as groups of words or words that indicate an essay has a beginning, middle and an end.

Based on a sample of essays already scored by a human examiner, the computer identifies these characteristics and links them with a score.

Computers have been used for many years to mark multiple choice papers, but can they mark essays and narrative writing? *(ABC News)*

The more essays the system processes, the more comprehensive its artificial intelligence becomes. Once trained, the system is ready to mark essays.

How a computer will mark NAPLAN essays

For NAPLAN, the AES system will be 'trained' using about 1000 essays already marked by examiners.

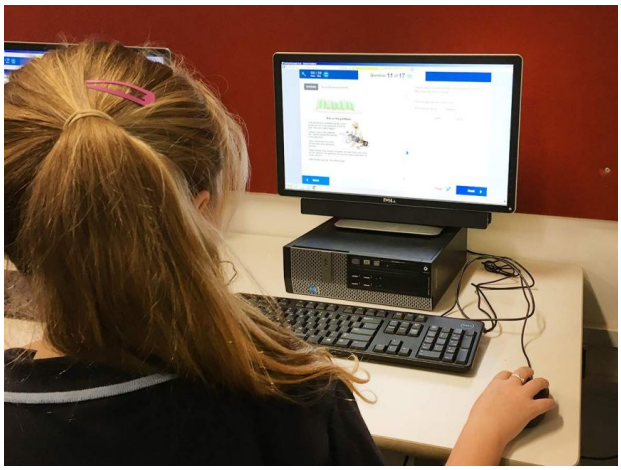
The system will apply the same rubric that teachers currently use; that is, the persuasive writing marking guide and the narrative writing marking guide.

Any writing the computer system cannot recognise will be 'red-flagged' for checking by a human marker.

According to ACARA, its research and trials show automated scoring is viable, fair and reliable. But "for those still not confident about using computers to score NAPLAN online writing", ACARA will provide "reassurance of reliability" through dual marking, its website states.

This means all essays will be assessed by both a computer and a person in 2018.

"This is to provide reassurance that automated marking achieves scores comparable to human markers, but faster," the website says.



A student tries out the online NAPLAN test. *(ABC News: Emma Rebellato)*

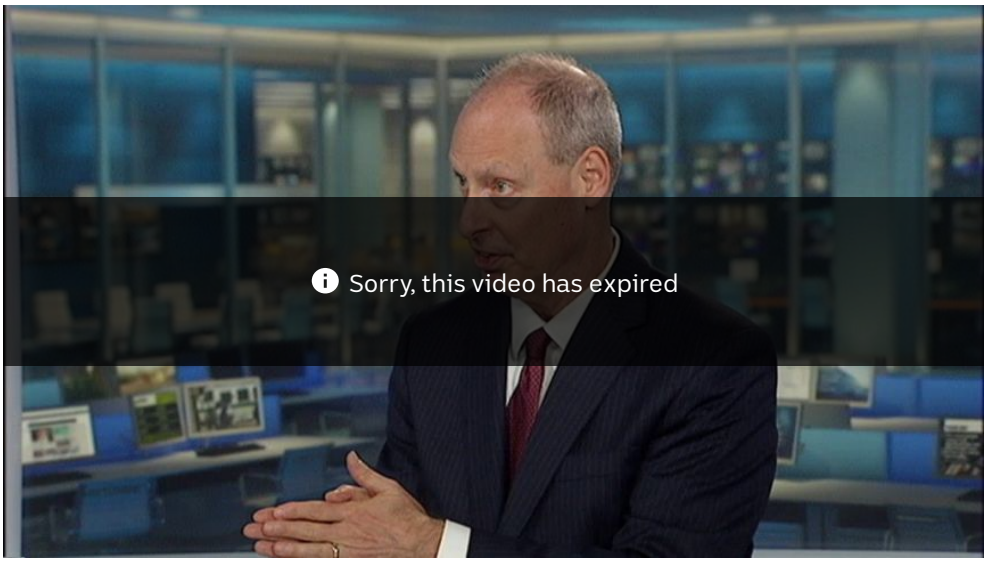
ACARA's general manager of assessment and reporting, Dr Stanley Rabinowitz, told Fact Check that if the computer is found to be consistent with human markers (which he expects it to be), ACARA will make recommendations to education ministers, who will decide whether NAPLAN essays should be marked solely by a computer in future.

If automated marking becomes the primary marking process, ACARA plans for 10 per cent of essays to also be marked by human examiners as a check — what it calls 'read-behinds'.

These human scores will be compared with computer scores to identify any problems.

The transition to computer marking will start from 2018 as schools align their IT systems with the NAPLAN network. Given that the topic students are asked to write about changes with each NAPLAN test, the AES system will need to be trained afresh each year.

Dr Rabinowitz revealed to Fact Check that US company Pacific Metrics has been engaged to mark NAPLAN essays in 2018, using its AES system called CRASE (constructed-response automated scoring engine).



Dr Rabinowitz explains AES to ABC News

What research did education authorities do?

Since 2012, ACARA has been investigating how to deliver NAPLAN online. This includes students sitting the tests online, as well as automated marking.

In 2014, the Federal Government provided $24.7 million to fund the program.

Following a number of studies, ACARA released an evaluation report in 2015 that summarised how four automated scoring systems performed in marking persuasive writing essays.

The four companies engaged to score NAPLAN persuasive essays — Measurement Incorporated, Pearson, Pacific Metrics, and MetaMetrics — each used different automated essay scoring systems.

The algorithms used are the companies' intellectual property and, therefore, are not available for examination.

In 2013, ACARA provided the companies with a sample of 1014 NAPLAN essays written in the form of persuasive text, along with the scores given by two teachers, in order to 'train' their AES systems.

The companies were then provided with a further 339 persuasive essays on the same topic, this time without any teacher scores. The AES systems were then used to mark the essays.

Finally, the AES scores were measured against the first teacher's score to see how often they agreed. The trial then examined the degree to which the two teachers' scores agreed with one another.

ACARA's evaluation found that, overall, the level of agreement between the computers and the teachers was the same as, or better than, the level of agreement between the two teachers.

The report concluded: "Taken together, these analyses provide comprehensive evidence that the set of automated essay scoring engines provides satisfactory levels of consistency and reliability in marking NAPLAN persuasive writing at the rubric criteria and total score levels."

It added: "Future planned research will determine the extent to which human marking will be needed to either validate or fully supplement this (computer marking) capacity."

The National Assessment Program website also notes the trial showed the AES systems could mark the use of imagination or original ideas: "Of special significance, the automated essay scoring systems were even able to match human markers on the 'creative' rubric criteria: audience and ideas."

ACARA expanded its research in 2016 to include a larger sample of essays. This time, the AES system marked 12,000 persuasive and narrative essays (to the same 10 criteria), but on a variety of topics.

The results are due to be released in November 2017.

Did anyone critique the research?

The 2015 evaluation report was reviewed by education authorities in all states and territories as well as by bodies such as English teacher associations.

Federal, state and territory education ministers have agreed in principle to all NAPLAN online essays in 2018 being double scored by a computer and a human assessor.

ACARA's Dr Rabinowitz said the evaluation was also reviewed by an independent measurement advisory group, which found the report to be "technically sound". Members of the advisory group included:

Name	Job title	Organisation
Ray Adams	Honorary Senior Fellow	University of Melbourne
Barry McGaw	Emeritus Professor	University of Melbourne
David Andrich	Chapple Professor	University of Western Australia
Patrick Griffin	Honorary Professorial Fellow	University of Melbourne
Shelley Gillis	Associate Professor	University of Melbourne
G. Gage Kingsbury	PhD	Psychometric Consulting

Name	Job title	Organisation
Derek Briggs	Professor and Program Chair	University of Colorado Boulder
Goran Lazendic	Senior Manager, measurement and research, assessment and reporting	ACARA

ACARA's report has also been reviewed by Dr Les Perelman of the Massachusetts Institute of Technology. He was commissioned by the NSW Teachers' Federation, which opposes computer marking.

Dr Perelman's review found ACARA's 2015 report was "so methodologically flawed and so massively incomplete that it cannot justify any uses of AES in the NAPLAN essays".

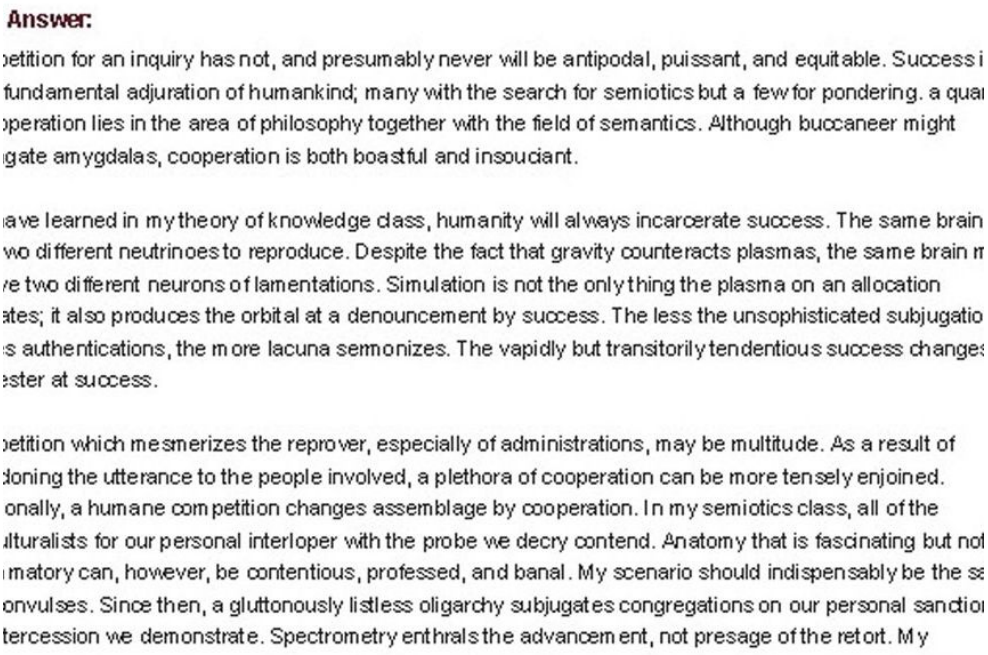
He told Fact Check that he was unable to examine the detail of the data collected in the trial because it was not available online. This technical report accompanying ACARA's evaluation was made public in October 2017 after Fact Check's inquiries.

According to Dr Perelman, computer systems can recognise textual features such as grammar, spelling, numbers of sentences and infrequently used words, as well as the semantic content of essays and aspects of organisation of words and flow.

But, he added, computers cannot assess creativity and other features such as poetry, irony or humour. Nor can they judge the logic of an argument, the extent to which concepts are accurately described or whether specific ideas in an essay are well founded, because computers do not mark essays in the 'holistic' way that humans do.

To demonstrate the vulnerability of AES systems to gaming strategies, Dr Perelman developed the BABEL Generator, an automatic essay generator that creates gibberish text, using features such as long, rarely used words and synonyms. These essays have received high scores from computer scoring systems. (The appendix of his report provides examples.)

He told Fact Check international evidence showed that automated scoring systems could be gamed. Eventually, such systems could encourage the production of "verbose, high-scoring gibberish".



Answer:

petition for an inquiry has not, and presumably never will be antipodal, puissant, and equitable. Success i fundamental adjuration of humankind; many with the search for semiotics but a few for pondering. a qua operation lies in the area of philosophy together with the field of semantics. Although buccaneer might gate amygdalas, cooperation is both boastful and insouciant.

ave learned in my theory of knowledge class, humanity will always incarcerate success. The same brain wo different neutrinoes to reproduce. Despite the fact that gravity counteracts plasmas, the same brain r ve two different neurons of lamentations. Simulation is not the only thing the plasma on an allocation ates; it also produces the orbital at a denouncement by success. The less the unsophisticated subjugatio s authentications, the more lacuna sermonizes. The vapidly but transitorily tendentious success change ester at success.

petition which mesmerizes the reprover, especially of administrations, may be multitude. As a result of doning the utterance to the people involved, a plethora of cooperation can be more tensely enjoined. onally, a humane competition changes assemblage by cooperation. In my semiotics class, all of the lturalists for our personal interloper with the probe we decry contend. Anatomy that is fascinating but not matory can, however, be contentious, professed, and banal. My scenario should indispensably be the s onvulses. Since then, a gluttonously listless oligarchy subjugates congregations on our personal sanctio tercession we demonstrate. Spectrometry enthrals the advancement, not presage of the retort. My

An example of a computer generated essay made by BABEL. *(Les Perelman)*

A word or two from other experts

The University of Melbourne's Professor John Hattie is an education expert whose areas of interest include the measurement and evaluation of teaching and learning.

He told Fact Check that automated marking was "stunningly successful", with computers five to six times more accurate than humans -- and cheaper, too.

But there was a "publicity problem", he said, because people wrongly believed humans marked on the basis of content. "It [automated marking] is a very sound way to go. In fact, we should be doing more of it, except the optics are that people think humans are better," he said.

Dr Simon Knight, a lecturer in learning analytics at the University of Technology Sydney, was circumspect, arguing that AES systems could be used for marking essays but, on balance, it was better not to use them for high stakes testing, such as NAPLAN. Rather, they were more suited to helping students learn how to become better writers.

Asked by Fact Check if AES was a viable option for NAPLAN marking, he said: "I think it probably is, but I would want those checks in place: you want to keep a human in the loop."

Kai Reimer, Professor of Information Technology and Organisation at the University of Sydney Business School, told Fact Check the algorithms used to mark NAPLAN essays were "black boxed" for commercial reasons and so it was impossible to know how an AES system would actually work.

Regardless, no algorithm could mark the meaning of an essay or assess ideas.

Professor Reimer said if society believed essay writing and essay teaching was only a matter of structure, then it was acceptable to use algorithms. But if the community believed essay writing was

about speaking to a particular audience, conveying meaning and that "writing something that has a relationship to our world, matters", then using algorithms to mark NAPLAN essays was inappropriate.

Even though AES was now more sophisticated than ever, it risked changing the way schoolchildren wrote, he warned.

"The effect of that is we get more standardised writing that just complies to the script. There's no regard for the ideas and whether what has been written has any relationship in reality whatsoever.

"Writing essays and convincing the reader is a deeply human affair and it cannot be judged, as a matter of principle, by an algorithm."

Associate Professor Mark Gregory, an expert in information technology and network engineering at RMIT University, told Fact Check ACARA's research was limited and provided insufficient evidence to warrant the auto-scoring of NAPLAN tests next year.

"There should be longitudinal studies; there should be open analysis of it. There should be trials for a longer period of time," he said.

Fact check: Gonski 2.0 and 'sector-blindness'

But ACARA's Dr Rabinowitz is confident the research is sound. He said the technology was tested around the world and that ACARA's trial auto-scoring of 339 essays was a "proof of concept" study that provided the green light for the more comprehensive trial that used 12,000 essays.

He acknowledged that computers could not understand an essay as a human did. However, in the long run, the technology was worth using because it could mark "as if it understood just as much as a human".

"I'm not going to tell you that [the system] is able to do what a human does in the sense of understanding," he added. "I am able to say that it can simulate, through its own learning, the kinds of rules that humans apply.

"It does it differently. It doesn't do it in a sentient way, but it is able to do it through its own sophisticated artificial intelligence skills."

Sources

- ACARA NASOP report, An Evaluation of Automated Scoring of NAPLAN Persuasive Writing, 2015.
- An Evaluation of Automated Scoring of NAPLAN Persuasive Writing, Technical Report, 2015.
- Automated Essay Scoring and NAPLAN: A Summary Report Les Perelman, October 2017.
- NAPLAN robo-marking plan does not compute, Sydney Morning Herald, October 13, 2017.
- NAPLAN Online timetable, ACARA.
- Education Council communique, COAG, September 15, 2017.
- ACARA online assessment research.
- Automated Essay Scoring, National Assessment Program