Automated Essay Scoring (AES) Innovation Trial 2021 Summary

NCEA Online Programme

NZQA
NEW ZEALAND QUALIFICATIONS AUTHORITY
MANA TOHU MĀTAURANGA O AOTEAROA
QUALIFY FOR THE FUTURE WORLD
KIA NOHO TAKATŪ KI TŌ ĀMUA AO!

2021 Innovation Trial Summary: Automated Essay Scoring (AES)

Background

In 2019 NZQA engaged the University of Alberta – Centre for Research, Applied Measurement and Evaluation (CRAME) to run a trial of automated essay scoring (AES). CRAME taught an AES scoring system how to mark a selection of anonymised NCEA digital exam scripts for a selection of English and History standards. CRAME reported to NZQA on their results and findings in June 2020.

While this trial helped NZQA understand the practical implications of implementing an AES, we wanted to know more. In 2021 we completed a second AES trial with additional vendors.

Trial overview

There was uncertainty whether some of the issues identified in the 2020 trial were related to CRAME's AES system, or to AES technology in general. By trialling with additional vendors we sought to identify the cause of the issues and provide further opportunities to broaden NZQA's knowledge on AES technologies and consider how they could be applied in the NCEA context.

Objectives

The 2021 AES trial aimed to further investigate the research objectives of the 2020 trial by:

- engaging with a wider pool of AES vendors to design and complete an AES trial using NCEA 2020 exam responses
- further exploring how suitable the automated scoring of essay questions is to the NCEA context
- broadening our understanding of the indicative effort required to establish an automated marking capability, including the skills, lead-time, and appropriate numbers of scripts for machine learning
- increasing our knowledge about potential opportunities where existing scoring processes can be supplemented by AES technologies
- exploring, when scripts have been marked, what the sources of difference between human and automated marking could be and how automated marking could contribute to the overall quality assurance of marking
- understanding the number of responses required to achieve an acceptance level of accuracy
- exploring how AES affects the time required to mark and quality assure marking of NCEA exams.

Vendor selection

After an open Request for Proposals (RFP) process, due diligence and the signing of non-disclosure agreements, two vendors were contracted to perform AES trials - Rembiont Pty Ltd, and New Data Solutions Pty Ltd in partnership with data specialist Vantage.

Preparing for and conducting the scoring

We found there was a limited number of standards with a big enough sample to meet the vendors' requirements for scoring. This constrained the possible testing.

In total, 2,000 responses were automatically scored by both vendors, substantially fewer than the 31,103 responses scored during the 2020 AES trial with CRAME.

A selection of anonymised digital exam scripts was provided to both vendors, along with the associated assessment schedules and exemplar scripts.

The vendors took different approaches to scoring:

- New Data Solutions, in partnership with Vantage, used their proprietary artificial intelligence powered IntelliMetric® system to conduct the automated scoring following the training of their scoring model.
- Rembiont's AEG Engine used an algorithmic approach, underpinned by information theory principles, to conduct the automated scoring, which didn't require training (in contrast to deep learning artificial intelligence approaches).

Findings

Both vendors provided detailed scoring, analysis and reporting to present their findings. More information is available in their reports:
AES Trial 2021 – Final Report – Rembiont.pdf
AES Trial 2021 – Final Report – New Data Solutions.pdf

Lessons learned

The following lessons were learned during this project:

- Schedule the key milestones so that there is adequate time to receive results from the vendor and decide whether to go ahead with events such as workshops before giving participants (in this case markers) markers notice of the event. Workshops involving school staff should be scheduled for the school holidays.
- Ensure responses supplied to the vendors for marking are the correct ones. In this trial the responses should have been those that were remarked in the previous trial. We gave the vendor a selection of responses and would have needed to ask NZQA's Data & Data Analysis team to match the selection against the set remarked during the 2020 trial to understand how many had been remarked. If the number was not sufficient, we would have had to reconduct the remarking exercise.
- Seek an early decision on budget expectations. We spent a significant amount of time looking at how we could reduce the number of responses to be scored in order to reduce the AES vendor spend.
- Ensure we have adequate responses across all achievement levels for an AES system to be trained and for scoring. This trial has given us an understanding of the number of responses required to give confidence in the results.
- Data quality is reliant on everyone involved in marking following the process and rules. New Data Solutions found responses that didn't appear to meet the criteria for an achievement level. Our investigation found that the responses shouldn't have been included in the data set.
- One vendor's AES system was not ideally suited for use with NZQA History assessments.

Recommendation

It was recommended that no further investigation into AES be undertaken until there is greater uptake of digital exams with questions that require an essay response. A larger dataset would enable greater confidence in the results from any subsequent AES trial.