2018

Why Can't it Mark this one? A Qualitative Analysis of Student Writing Rejected by an Automated Essay Scoring System

Nathanael Reinertsen, *Australian Council for Educational Research (ACER)*

Why Can't it Mark this one? A Qualitative Analysis of Student Writing Rejected by an Automated Essay Scoring System

Abstract

The difference in how humans read and how Automatic Essay Scoring (AES) systems process written language leads to a situation where a portion of student responses will be comprehensible to human markers, while being unable to be parsed by AES systems. This paper examines a number of pieces of student writing that were marked by trained human markers, but subsequently rejected by an AES system during the development of a scoring model for the *eWrite* online writing assessment that is offered by The Australian Council for Educational Research. The features of these 'unscoreable' responses are examined through a qualitative analysis. The paper reports on the features common to a number of the rejected scripts, and considers the appropriateness of the computer-generated error codes as descriptors of the writing. Finally, it considers the implications of the results for teachers using AES in assessing writing.

Keywords: automated marking errors, assessing writing, automated essay scoring

Why Can't it Mark this one? A Qualitative Analysis of Student Writing Rejected by an Automated Essay Scoring System

Automated scoring of student writing is increasingly used in a variety of high-stakes tests across the world. Its proposed use in a large-scale national assessment program in Australia has recently been a contentious topic and the proposal was abandoned as a result (Robinson, 2018). The public debate has not allayed, and perhaps has even increased, suspicion about how Automated Essay Scoring (AES) systems score. Of greater importance to teachers, though, is that the use of AES in writing assessments is not limited to high-stakes or large-scale assessments; it is also used in assessments designed for classroom-based assessment. There is little evidence available as to how widespread the uptake of automatically scored writing assessments is in Australian schools, but well over 200,000 pieces of student writing has been scored by just the one AES system this paper focusses on in the last four years (*see* Table 1), and it is not at all the only automatically-scored writing assessment available.

With increasing awareness of AES in Australia, and its current use in Australian classrooms, it is important that teachers know about and understand the strengths and weaknesses of AES generally, but also the strengths and weaknesses of the specific assessments available to them.

One of the aspects of AES that receives little attention is what happens to student writing that cannot be parsed by the AES system. The difference in how humans read and how AES systems process written language leads to a situation where a portion of student responses will be comprehensible to human markers, but unable to be processed by scoring systems. It is this difference in what can and cannot be scored that this paper sought to explore through the examination of 23 pieces of de-identified student writing that were marked by humans, and subsequently rejected by an AES system during a development process.

It almost doesn't bear repeating that writing is an act of communication between humans, and that the act of writing is an attempt to communicate ideas to a reader. The reason for restating the obvious here, is that it is because of this fundamental communicative intent that it is rare for a human marker not to able to interpret at least some part of a student's writing, and thus it is rare for a human to be completely unable to assign a score based on a judgment of the writing's quality.

It might be at this point that a question arises about whether the rejection of scripts by the AES system is simply due to students' typing skills or the computer system not recognising badly misspelled words. As you will see later, there are indeed scripts that are rejected because a computer system does not recognise enough words, but there are also scripts which received high scores from human markers that were rejected by the AES being examined. These scripts in particular beg for more investigation. Observations about their individual qualities and the qualities they share with other rejected writing, can shed light on some of the limitations of this particular AES system when it comes to judging student work, and it may contribute to discussion on the broader issues around using AES in assessing student writing.

Background

Unscoreable scripts

The limitations of AES have received attention in published research (e.g. Deane, 2013; McCurry, 2010a, 2010b; Perelman 2014), in addition to body of research that has examined whether or not AES is a fair, reliable, appropriate, and/or valid assessment method (*see* Bennet & Zhang, 2016; Shermis & Burstein, 2013). However, there has been comparatively little reported about the particular situation of what happens when AES systems cannot parse the writing submitted to them, and even less consideration about whether it is only the poorest writing that is rejected by AES systems. A discussion of the rate and reasons for rejection of scripts is often absent from research reports, validity arguments and reliability studies.

Yet there is an opportunity to use rejected scripts in the evaluation of AES systems. In the first place, the proportion of scripts that will have to be manually marked is a factor in the efficiency of the assessment. In the second place, it can be used to more deeply investigate the agreement between the AES and human raters. Evaluating AES systems on statistical agreement rates with human markers is a practice that has been called the 'gold standard' of AES validation (Powers, Escoffery & Duchnowski, 2015). What has not been done often is to examine the scripts on which the AES system and the human raters had a difference of 'opinion', as much as an artificial intelligence can be said to have one, to investigate whether human raters value features of the writing differently.

So, unscoreable scripts are worthy of further investigation on two grounds. Firstly, they provide information on the operational efficiency of an AES system when evaluating its merits for a particular application. Secondly, it may offer insight into the differences between the reading practices of human markers and the textual analysis process of AES algorithms.

eWrite

The assessment from which the data for this analysis was collected is *eWrite*, an online writing assessment offered by ACER since 2014. The assessment is intended to be a classroom assessment, recommended for grades five to eight that provides data to a teacher for the purpose of informing the teacher's own practice. The teacher selects one of the writing tasks developed by ACER and made available through the online interface, for their class to respond to. There are four available genres of prompt: narrative, persuasive, descriptive or report. The students complete the assessment online with a 25 minute time limit.

eWrite's AES system uses Vantage Learning's IntelliMetric® scoring engine. The scoring models were developed for ACER by Vantage Learning from corpuses of student work sampled from ACER's trials of the writing prompts. Before being accepted for use and made available to clients, there is an evaluation of the reliability of the scoring model. During the development of the scoring model, Vantage Learning withholds 50 scripts from the supplied sample of student writing and has these marked by the system after the scoring model has been developed. The scores for these 50 scripts are compared to the scores the scripts were assigned by trained human raters, and the results of this comparison are reported to ACER for the purpose of evaluating whether the scoring model is fit for purpose.

The AES system is able to almost instantly score the student work upon its submission, and the scores are made available to the teacher through the online system immediately after the scoring is complete – except for those cases where the AES is unable to mark the student's work. In such cases, the reports for those students are blank, and the teacher or school has the option to request for the scripts to be marked manually by a trained marker, or to attempt to apply the marking guide provided in the test's documentation.

The *eWrite* marking guide, used by human markers only, is an analytical marking guide with a varying number of criteria depending on the genre. For example, the persuasive writing task is marked on nine criteria, with varying numbers of score points, and the total available number of marks is 28. AES systems do not apply marking guides, but the reports returned to teachers contain scores labelled with the same criteria, and the maximum number of score points in each criterion and in total is the same.

Unscoreable scripts and eWrite

From 2014 to the end of 2017, there were 230,845 scripts submitted to *eWrite* through ACER's Online Assessment and Reporting System. Table 1 displays the number of scripts that have been rejected by the AES system in the last four years. As can be seen, approximately 6% of all the student writing has been unable to be marked by the AES

system, although the annual rate varies from 4.8% to 7.3%. In other words, if a teacher has a class of thirty students who sit an *eWrite* task, it is likely that one or two students will have no scores returned by *eWrite*.

Table 1 Automated Scoring Model rejection rates 2014 to 2017

Scripts	2014	2015	2016	2017	Total
Total (n)	22,728	31,278	84,538	92,301	230,845
Unscoreable (n)	1357	1855	4048	6741	14001
Unscoreable as a percentage of total (%)	5.97	5.93	4.79	7.30	6.06

When the *eWrite* AES system encounters a script it cannot process, it records one of a number of possible error codes, which are displayed in Table 2. If one were to judge only by the names of the errors, it may appear that the AES system is incapable only of marking error-prone writing: scripts full of words misspelled in such a way as to be unrecognisable, or having very poor grammar. However, this is not always the case.

Table 2 Errors codes generated by the eWrite automated scoring system

Error	Definition
Off-topic	essay does not contain a minimum number of words from the prompt-specific lexicon built during creation of the scoring model
Bad Syntax	insufficient punctuation/too many run-on sentences; syntax errors which prevent understanding
Bad Vocabulary	spelling is overwhelmingly poor or essay is written in a foreign language
Repetitious	text or sentence structure is repeated

A new prompt was trialled at the beginning of 2016, and 531 marked scripts from that trial were sent to Vantage Learning for the purpose of developing a scoring model for that prompt. Of those 531 scored scripts, 23 were rejected by Intellimetric® with error codes. Table 3 displays the frequencies of the error codes generated for these 23 unscoreable scripts, alongside the total raw scores that were assigned by a human rater to those scripts.

Table 3 Error codes for unscoreable scripts in a training sample from 2016, with total scores assigned by human raters

AES Error code(s)	Frequency	Total Scores Assigned by Human Raters
Bad Syntax	12	2, 3, 4, 5, 6, 11, 12, 14, 15, 15, 16, 17
Off-topic	6	0, 1, 2, 8, 17, 17
Bad Vocabulary	1	3

Repetitious	1	15
Bad Syntax; Off-topic;	2	0, 0
Bad Syntax; Bad Vocabulary; Off-topic	1	0

(n=531)

There are a several aspects of these data deserving of comment at this point. Firstly, three scripts were assigned multiple error codes. It may be reassuring that these scripts had all received total raw scores of zero from the human markers, which can be taken to indicate that these pieces were in significant need of revision in some way. The second point deserving of comment is that six scripts were coded as off-topic by Intellimetric®, but that those scripts had been assigned a range of scores between 0 and 17 by human markers, and the higher scores in particular raise the question of why they were identified as being off-topic by Intellimetric®. The third point is the wide range of scores for the scripts that were assigned a 'Bad Syntax' code. Clearly there is a wide range of quality in the writing that has been identified as unscoreable and there are questions to ask about why scripts that appear to be of moderately good quality have been rejected.

Aims

This paper intended to address two aspects of the 23 scripts that were rejected by the *eWrite* AES system during the development of a new persuasive writing task in early 2016. In the first place, it aimed to identify writing features that may be shared between several of the pieces of writing. Secondly, it intended to identify whether the computer-generated error codes are appropriate descriptions of the writing. These questions provide teachers who use *eWrite* insights into the reasons a student's writing may be rejected, and there is a possibility that these insights might be generalisable to other automatically-scored assessments, though any generalisations would require further research. Deeper understanding of the limitations of this AES system, such as what it can and cannot score, will hopefully lead to better evaluations of when it may be an appropriate assessment method.

The above aims of this research can be expressed as two research questions:

1. What writing features are shared between scripts rejected by the *eWrite* AES?
2. Are the computer-generated error codes appropriate descriptors of the rejected writing?

Methods

As referred to previously, a new persuasive writing prompt was developed, trialled and added to the *eWrite* assessment in 2016. The task is a writing prompt that asks students to write to convince a reader to accept their opinions about the value of books. The participants in the trial were 1050 students from eleven schools, comprising both independent and state schools from Victoria and Western Australia. The age of students was not collected, but the trial sample included a range of grades from Year 4 to Year 10. The writing was collected under ACER's Online Assessment and Reporting System (OARS) Terms and Conditions, which explicitly allow for de-identified data collected through OARS to be used for research purposes, both by ACER and by third parties.

A training sample of 531 scripts from the trial sample was prepared for submission to Vantage Learning to develop the automated scoring model. The scripts were selected to approximate a normal distribution of scores across the entire score-range. Of those 531 scripts, 23 were rejected during the development process by the AES system. These 23 scripts were located in the test data and extracted, forming the sample for analysis in this paper.

The analysis of the scripts followed the Interactive Model of analysis as described by Miles, Huberman & Saldaña (2014). The model outlines a process that comprises four components: data collection, data condensation, data display, and drawing or verifying conclusions from the data. These components in the Interactive Model are considered as part of an iterative, concurrent process.

The author supervised the marking of the trial assessment which provided the data for this research. As such, he was familiar with the assessment task, the marking guide, and had read and scored a range of student responses. To investigate the rejected scripts in more detail, he read through each of the 23 scripts, and recorded a general comment without viewing the script's score or error code. He focussed rather on describing observations about apparent strengths and weaknesses in the scripts. This process was undertaken in order to condense the information in the data in preparation for coding. The comments were then read over, again by the author, in conjunction with the scores and error code. Codes were created organically and iteratively with increasing levels of abstraction, with names for the codes being selected to represent commonly observed features of the scripts, such as 'missing punctuation'. After coding, a simple visual display was constructed in order to identify common qualities of the scripts. The visual display was a table, where each row was a script, and each column a code. Where the code was attributed to a script, the corresponding cell

was shaded, using a different colour for each code. From the comparison of common features, the scripts were grouped into four broad categories that are described below.

Results

The scripts were divided into four categories according to their score attribute and the codes assigned to them, as described below. The categories arose during the coding process, and are used to abstractly represent scripts that were judged to share some qualities.

Category one

This category comprised three scripts that were rejected by the AES system with multiple error codes. All three were tagged with both 'Bad Syntax' and 'Off-topic', and one was additionally tagged as 'Bad Vocabulary'. All three scripts had been scored 0 by human raters, and upon examination the reasons were apparent. Take as an example Writing Sample (WS) 307, which was the script tagged with all three error codes: "i LIKE DSADQDADADSSDADSDADADADADADADADADAADAD" (WS307)
The limitations of the script are evident enough that not much discussion is warranted: the three codes are wholly appropriate.

However, WS 2554 was a little different. It was tagged as 'Bad Syntax' and 'Off-topic', and reads, "books are important because they do't use power they". An insufficient response, certainly, but a human scorer would probably not recognise this script as being off-topic – there is an idea here about books. It is undeveloped, very short, and unfinished, but the student has attempted to engage with the writing prompt. A score in the lowest category is warranted, but it is not an off-topic script. The 'Bad Syntax' error code appears to be suitable, although making such a judgment based on so little writing is difficult.

Category two

All the scripts in the score range of 0-6 (n=13) were coded as 'Lacking punctuation', and this classification was used to define the boundary between categories two and three. Because four of those thirteen scripts were classified into category one because of multiple error codes, there are nine scripts in category two. Writing sample 39 is illustrative of scripts in this category:

> i think reading is more useful than finding stuff on a computer could either be fake or not a very good source which is why i would rather a book because it has to go through a publisher where as books on a computer can be posted by anyone and might not have true facts in it if its a fiction book and might be written by an ameteur where as books are written by professionals if its for eduactation purposes (WS39)

The writer of this script has an opinion, and reasons for it. The writing is able to justify the writer's opinion to a limited extent, and the communicative intent is clear. There are a handful of spelling errors, and the major flaws of the writing are the absence of punctuation and the fact that the whole text is a single, run-on sentence. The AES system rejected this particular text with a 'Bad Syntax' code, which does seem appropriate. Six of the nine scripts in this category that were coded as 'lacking punctuation' received the same computer-generated error code. This suggests there may be a relationship between lack of punctuation and the AES system being unable to parse the writing.

One of the more remarkable scripts in this category was rejected for 'Bad Vocabulary'. It reads, in part:

> When I ferst read a book I thort that it was boring aswell but when I got to the midole of the book
>
> I fand it realy intresting. I think books are good because thay halp you with your speling and comprerhenchen.
>
> My thorts on books are that you are use you amagenachon and it can cume you down when you are strest. (WS723)

Some of the words in this passage are spelled in a way that suggests a reliance on phonetic approximation. The words are interpretable (for a human) despite the flawed spelling. One can infer from the computer-generated error code, though, that misspelling to this extent interferes with the computer's ability to analyse the text.

There were two scripts in this category that were labelled 'Off-topic' by the AES system. Both were very short scripts, but each contains a statement about books. For example, WS 1871 begins with the statement: 'Books have been used for over a thousand years and they seem to have no end in sight.' This appears to be an inappropriate error code and the same is true of the other script with this tag in this category. Taken with the other off-topic error code in the previous category (WS 2554), it appears there may be a relationship between very short scripts and the AES assigning an off-topic error code.

Category three

This category was defined as the scripts which were generally correct in terms of surface language conventions (punctuation, spelling, etc.) but that were not coded as 'developed' like the scripts in category four. Using this definition, this category comprises seven scripts with a score range of 8-15. In the computer-generated error codes, there are five 'Major syntax error', one 'Repetitious' and one 'Off-topic'. The extract below is indicative of the general standard of scripts in this category:

> First of all you should read books to improve your comprehension because you use your comprhension from any age and you use it in all subjets. If you are good at coprehention you can understand all hard questions that you are asked. If you can not read well you will struggle in highschool in all of your core subjects and you may find it harder to get a job when your an adult. (WS674)

It is a piece of writing that is generally correct in its spelling and punctuation, that expresses an opinion in generally correct language, but that does not fully develop its idea nor strongly convince the reader to accept the idea it presents. The computer-generated error code for WS674 is 'Bad Syntax'. It is difficult, not only in the extract above, but in the whole piece of writing, to identify what these major errors might be. There are certainly some misspellings, and the sentence structures are not elegant, but it is difficult to justify calling them major errors. In fact, it would probably be more appropriate to attribute the misspellings to carelessness rather than evidence of spelling ability as the misspelled words ('comprehension' and 'subjects') are each spelled correctly once in the paragraph in addition to the instances where they are misspelled.

The same error was generated for WS3900 which is a stronger piece of writing, though it too does not quite reach the level of a developed argument, and its logic, syntax and vocabulary exhibit some errors. An extract:

> Books teach us an understandable and easy way to learn about new topics that we have not heared before, they also make you think and take you on an amazing adventure and rollercoaster around the world. Books also cature for anyone and everyone as there a lot of different genres and different types of writing. (WS900)

Once again, to reject this as unscoreable for 'Bad Syntax' seems difficult to explain – there are errors, but the errors do not amount to the script being incomprehensible for a reader. One other script in this category received the same error code and there is a similar difficulty in justifying that label. In three other scripts, the possible relationship observed earlier with regards to punctuation errors and the 'Bad Syntax' code would help to explain the reason for those scripts being labelled with that error code: they did not lack punctuation, but there were frequent errors in punctuation.

There were two Category three scripts different to the others in terms of their error codes. One was deemed off-topic and the other repetitious. The off-topic script certainly bears discussion because the student has written a narrative response to the prompt, and the script is different to the majority of other responses for doing so. The narrative is about a boy who gets in trouble for coming home late because he loves books and lost track of time while

at the library. Such a narrative was clearly written as a response to aspects of the prompt, and could conceivably be an attempt to convince a reader to sympathise with the protagonist's opinion of books:

> "Jimmy have you any idea of what time it is?" scoulded his mother.
>
> Jimmy knew that this was one of those questions were she already knew the answer and he didn't understand why she asked.
>
> "why do you spend all that time at that stupid book thingie?" again with these questions Jimmy just didn't get it.
>
> "because i love books" retorted Jimmy using every strain of courage in his body. He ran to his room without dinner and cried for a while. (WS 3988)

There is creativity here, and a self-reflectiveness in the observation of the mother's pointless questions. This is writing that it is easy to imagine English teachers encouraging. Yet, it certainly is different to most responses, and that uncommon approach to the prompt appears to have been identified and rejected by the AES system. However, labelling such a script 'off-topic' would likely be a contentious point among human markers: at the least it would be a topic for discussion and clarification.

The 'Repetitious' error code that was received by one script in this category does seem to be an appropriate descriptor of the writing. It was a persuasive text with one main idea, and each paragraph was a minor variation of the idea, and some words and phrases are repeated a number of times across the script.

Category four

This category comprises the three scripts that were coded as 'developed', and all three were the top-scoring scripts in the sample with scores of 17. One suffers from frequent misspellings and received the 'Bad Syntax' error code, though why that code rather than 'Bad Vocabulary' is unclear.

Two scripts were narratives and they both received the 'Off-topic' error code. One of them begins with a dream in which the protagonist is delivering a speech to her class:

> "Books are a way to communicate stories and important information. The first great civilisations created books. We are ancestors of these great civilisations, why do we insist on changing thesw simple ideas. They survived for thousands of years, yet all we have done is develop our instrument, our tools, we have not developed our minds. Books allow us to do this. We need books. They are an important part of our social lives, we reccommend books to friends and peers, then we can talk about the events in the books. If we don't have books, we don't have education. " I finshed My debate

> with a bow and a smile. Everyone in the audience applauded me. Then the principle walked in, everyone went silent. (WS3964)

The narrative proceeds to change into one where the protagonist's brother gets magically sucked into the pages of a book, and ends with the protagonist going in after him. Is this piece of student writing off-topic? It is certainly dissimilar to the majority of responses to the prompt, but the opening paragraph is a better persuasive piece than some of the category one and two scripts, and does directly engage with the writing prompt.

The other category four, 'Off-topic' narrative is a dystopian vision of computers rising up to take over the world. They are defeated when a messianic figure, named Chris, recreates a book. The people of that world revere it as a holy object and then overthrow their robot overlords. However:

> The people had spent so long without computers they took books for granted. This made them easy pray for the minions of the computers. Soon Chris was defeated and his followers all killed or converted. The computers then went on to make sure that humans could never belive in books again. (WS 3978)

Once again we are faced with the question of whether this is an off-topic response. It is very different to the majority of scripts, but it is centred on books and their value. Whether its purpose is to convince a reader to accept an opinion, though, is far less clear, and it would not be too hard to imagine human markers disagreeing about whether this is on- or off-topic, in which case perhaps the AES rejecting such pieces could be seen as a feature, rather than a software 'bug'.

Discussion

This research aimed to answer two questions: whether there were common features among scripts rejected by the *eWrite* AES, and whether the error codes generated by the *eWrite* AES system for the rejected scripts were appropriate.

Common qualities of rejected student writing

There were a total of 23 scripts rejected by the *eWrite* AES system: twelve of those were coded as 'lacking punctuation', with a further two described in the initial commentary as missing some punctuation. Eleven of those fourteen scripts with punctuation problems were rejected by the AES system with the 'Bad Syntax' error code. This is too small a number of scripts from which to draw firm conclusions, but there is an indication that punctuation errors are a common factor that contributes to scripts being rejected by the *eWrite* AES system. This is, perhaps, unsurprising if one considers punctuation as being a

way for a writer to indicate the boundaries of words and sentences; automated analysis appears to struggle when such boundaries must be inferred rather than observed.

Appropriateness of error codes

There were two computer-generated error codes where the qualitative analysis indicated the codes may not be wholly appropriate. The most common error code across all categories was 'Bad Syntax', and while the code seemed appropriate in most cases, there were three cases where the AES error code does not appear to describe the writing. The sample is too small to draw firm conclusions, and while these error codes seem inappropriate in these particular instances, more research is needed to argue that the error codes are incorrect. For example, a quantitative analysis of a larger sample to investigate an error to length ratio might possibly explain why these scripts were rejected.

However, every instance of the 'Off-topic' error code was found to be hard to justify. In categories one and two, it was associated with very short pieces of writing. But even in the few words that were written there, the qualitative analysis found that the writing at least referenced key words in the writing prompt. More seriously, though, was the 'Off-topic' error code being applied to narrative responses to the writing task in categories three and four. All three narratives among the reject scripts were tagged as off-topic, but all three made reference to aspects of the writing task, though they arguably lacked in persuasive effect. The rejection of scripts that take an alternative or oblique approach to the writing task supports McCurry (2010a; 2010b) in suggesting that automated scoring is less likely to be able to deal with 'broad and open' writing, and is more appropriate for constrained tasks.

There is not enough of a basis to conclude that writing of the 'wrong' genre will always be rejected by the *eWrite* AES system, however there is an indication that it is likely to be labelled as off-topic and not scored by the AES system. This is in accordance with the literature that has found that the writing construct that is assessed by AES systems is restricted, mostly providing evidence about surface features of writing rather than features such as content, form and effectiveness (Condon, 2013; Deane, 2013). What is new about the research being reported in this paper is that these narrative pieces of writing were included in the training corpus, yet were rejected with an error *despite* the fact that the training corpus is explicitly intended to form the basis of the scoring system's development. This suggests that including a broader range of writing styles in the training corpus will not result in an AES system that is more able to score a broader range of types of responses. This raises questions about the development process that ought to be the subject of further research.

Whether the rejection of writing for being dissimilar to the majority of scripts used in the development of a scoring system is an acceptable feature of *eWrite* will depend upon the purpose a teacher or school has for administering the assessment. If the teacher assigns a persuasive task to their class for the purpose of assessing the students' persuasive writing, then rejecting writing that does not have an obvious persuasive intent might be useful as it would ensure the teacher has an opportunity to review the piece and make a professional judgment about the writing performance. However, if the purpose is to assess a broad writing construct, then rejecting any piece of writing that does not match the style or content of the training sample will likely lead to a higher rejection rate. This would significantly undermine one of the intended advantages of automatic scoring: providing scores quickly.

Conclusion

The process by which automated systems score student writing are unlike the processes used by human markers. This leads to a situation where some scripts, as seen in this research on the *eWrite* assessment, cannot be marked by an AES system despite being competent acts of communication between writer and reader. Rejecting scripts that have very poor punctuation or spelling is, perhaps, understandable. The majority of rejected scripts in this small sample featured missing or incorrect punctuation, suggesting that the AES system relies more heavily on this feature to process written language than does a human marker, and this is unsurprising given previous research (Condon, 2013; Deane, 2013).

The main implication for users of *eWrite* (or similar assessments) from this present research, is that some writing that receives moderate to high scores from human markers will be rejected as unscoreable and 'Off-topic' because it is dissimilar to the majority of the training corpus. Teachers and school administrators ought to bear this in mind when deciding whether or not to use an automatically scored assessment in a particular context, for a particular purpose. In the context of using AES in a classroom assessment, though, it is unlikely to be a serious issue.

The findings of this paper support a recommendation that teachers' professional judgment should be used in reviewing scores from the *eWrite* assessment alongside the scored scripts as well as closely examining the 'unscoreables', to ensure that the error codes and scores are appropriate, before using the scores as the basis for providing feedback to students. Because where a student's writing has some features that are dissimilar to the majority of responses, there is some likelihood that the scoring may not be accurate. This recommendation is likely to be good practice for any classroom-based, automatically-scored writing assessment.

References

Bennet, R.E, & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing: improving educational and psychological measurement.* New York: Routledge.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings?. *Assessing Writing 18*, 100-108. doi:10.1016/j.asw.2012.11.001

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing 18*, 7-24. doi:10.1016/j.asw.2012.10.002

McCurry, D. (2010a). The Machine Scoring of Writing. *English in Australia 45*(1), 47-52.

McCurry, D. (2010b). Can machine scoring deal with broad an open writing tests as well as human readers? *Assessing Writing 15*, 118-129.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (Third edition.). Thousand Oaks, CA: Sage Publications

Perelman, L. (2014). When 'the state of the art' is counting words. *Assessing Writing 21*, 104-111.

Powers, D., Escoffery, D., & Duchnowski, M. (2015). Validating Automated Essay Scoring: A (Modest) Refinement of the "Gold Standard". *Applied Measurement In Education, 28*(2), 130-142. doi:10.1080/08957347.2014.1002920

Robinson, N. (2018, January 29). NAPLAN: Robot marking of school tests scrapped by education ministers. *ABC News Online.* Retrieved from: http://www.abc.net.au/news/2018-01-29/push-to-have-robots-mark-naplan-tests-scrapped/9370318

Shermis, M. D., & Burstein, J. (Eds.) (2013). *The handbook of Automated Essay Evaluation: Current Applications and New Directions.* New York: Routledge.