MARC: a thought experiment in the morality of automated marking of English

Victoria Elliott

Oxford University Department of Education

velda.elliott@education.ox.ac.uk

MARC: a thought experiment in the morality of automated marking of English

Automated essay scoring programmes are becoming more common and more technically advanced. They provoke strong reactions from both their advocates and their detractors. Arguments tend to fall into two categories: technical and principled. This paper argues that since technical difficulties will be overcome with time, the debate ought to be held in terms of the principles. A thought experiment, based on a technically perfect Automated Essay Scorer, is proposed in order to explore the moral questions related to this topic, such as whether students deserve to have their work read by a human. It concludes that affect is an important component both of writing and of the debate, but that if the move to automated scoring stops being an 'all or nothing' debate, then many of the objections on principle will be obviated.

Keywords: automated essay scoring, essay marking, assessment.

Background

Computer programmes designed to mark essays, known as Automated Essay Scorers (AES), have been in the public eye for some years now, sparking the occasional news item when a speech by Churchill is fed into a programme and promptly assessed as 'below average' (BBC, 2009). AES are rapidly becoming a more realistic and sought-after prospect. They create algorithms to analyse text, based on training scripts previously marked by humans, and sort subsequent inputs into categories (grades). In April 2013 EdX, the non-profit organisation formed by a conglomerate of US universities to deliver MOOCs announced it would make its automated grading software available to educational establishments for their own use. Some greeted this as a major breakthrough, as previous programmes were severely protected as part of their hoped-for profitability. However, an online survey by the *Guardian*'s education site after the announcement found 87% of 295 respondents were against the idea of teachers using software to grade students' work (*Guardian*, 2013).

The issues surrounding AES can be divided into technical objections and objections on principle. A technical objections might be that AES use proxies, such as the 'length of word' and 'relative infrequency' instead of assessing the underlying dimension

‘sophistication of vocabulary’. Such objections are temporary at best. The identification of technical objections should be welcomed by advocates of AES, as the more quickly they are identified, the more quickly they will be overcome. In addition, while opponents are fighting the battle against automated essay scoring on these grounds, they are not fighting it on the fundamentals. The objections based on fundamental principles are less easy to overcome, and to some extent represent a point at which the two sides of this debate are simply unable to comprehend the position of the other.

For this reason I offer a thought experiment in the morality of AES. Technical objections to their use ought not to drive decisions about them: those objections will eventually be overcome. Ellis Batten Page projected the potential of AES in 1966 to ‘rescue the conscientious English teacher from his backbreaking burden’ (1966, p. 238); he subsequently abandoned practical experiments for twenty years, on the basis that the computing power available at that time would have made the cost prohibitive. It is on the basis of the principles of the matter that we ought to debate and make decisions about the place of AES in education, and do so before they are presented as a *fait accompli*. I will first discuss some of the prominent arguments about the use of AES, before turning to the thought experiment, the creation of ‘MARC’, and some ethical and moral issues about the potential use of AES.

The debate

One prime objection is that teachers will change the way they teach in order to give advantage to the features which AES reward. The ‘Professionals Against Machine Scoring Of Student Essays In High-Stakes Assessment’ website feared teachers will be:

> ‘coerced into teaching the writing traits that they know the machine will count– surface traits such as essay length, sentence length, trivial grammatical

> mistakes, mechanics, and topic-related vocabulary—and into *not* teaching the major traits of successful writing—elements such as accuracy, reasoning, organization, critical and creative thinking, and engagement with current knowledge' (Human Readers, 2013).

A cynic might point out that it is 'surface traits' that the introduction of the Key Stage Two Spelling and Grammar Test in England is designed to promote. An idealist might suggest that teachers will teach what they know to be the fundamentals of good writing, rather than whatever will enable students to game the test, but history suggests otherwise, when it comes to high stakes examinations.

One potential solution to the problem of washback is to use essay marking software as a way to allow teachers to set many more practice assignments, with instant feedback, not replacing human marked essays, but supplementing them. The US school system has historically a lesser emphasis than the UK on creative writing and a correspondingly greater one on argumentative and analytical essays, where such use might be more applicable. In the UK the use of AES might be most likely in terms of GCSE English essays, to provide a second marker in the same way that they are currently used in the GMAT test in the USA (Deane, 2013). As a second marker, AES should not lead to washback effects, since the primary marker would identify those gaming the system and penalise them.

Research has been conducted on ways in which AES algorithms may be gamed. Powers *et al.* (2001) challenged individuals to write essays which could trick ETS's proprietary *e-rater* software into wrongly scoring them. They found that it was relatively easy to do so, but considered that this produced avenues for improving the software, rather than reasons for abandoning it. As systems and their algorithms become ever more complex, the challenge of writing an essay that can game the system will become correspondingly harder, to the point where it will be more effort to do so than simply to learn to write the essay in the

first place (and many of the surface traits that would be required are still valuable points of learning, such as the acquisition of polysyllabic vocabulary). At some point the technology will reach a level of sophistication where a student who has managed to game the system will deserve an A for ingenuity.

There is another class of argument which suggests that AES are fundamentally flawed because they rely on proxies rather than the underlying dimensions we would wish to consider for valid assessment of English writing. Proxies are characteristics which an AES can identify and which it can use to 'predict' a text's outcome on an underlying dimension. These may be related, as in the case of *word length + relative infrequency → sophistication of vocabulary*. They may be completely unrelated, which is the danger, but completely unrelated proxies are unlikely to continue to predict performance throughout an entire set of, say, 100 training scripts (the word 'predict' relates to the way the AES tests its algorithm, 'predicting' a score, checking its 'true' score and then adapting its algorithm as appropriate). There is a difficulty in the potential for an opaque system: if it is using unrelated factors then there is the potential for mis-grading, and questions of validity are raised.

However, this argument assumes that we know that markers make valid decisions and are not using proxies to reach judgements quickly. Although we do know on what human markers are *meant* to be basing judgements (mark-schemes and Assessment Objectives), research into judgement and decision-making makes it clear that assessment is more than the sum of its parts (Sadler, 1989). We assume that the underlying dimension is being assessed, but too little is known about the thought processes of markers to be sure. In essays there is an added level of complexity: is clarity of written communication an integral part or merely an obfuscatory factor which is likely to be used as a proxy for quality of literary or linguistic analysis, as a more easily seen 'surface trait'?

Other false dichotomies have been proposed as differences between automated and human essay grading. One is that as an AES cannot identify ‘truthfulness’, then students will be trained in the art of ‘fabrication rather than the researching of supporting information’ (Human Readers, 2013). The idea that fictioneering is not prevalent in human marked essays is an interesting one: it presumes that markers have the full range of potential knowledge and are expected to spot created facts. My own research suggests that unless markers know something which directly contradicts a statement in an essay, if that statement appears plausible, it will be accepted (Elliott, 2011). In English writing one might argue the form matters rather than the facts, outside of the disciplines of literature and linguistics. It is already commonplace for GCSE students writing persuasive essays to make up statistics relating to smoking, or school uniform etc.

The debate over AES is frequently heated and emotive, and nowhere more so than in the application of automated scoring to famous texts. In 2009 the media reported an experiment in which a speech of Churchill’s, a passage from *Lord of the Flies* and some of Hemingway’s writing were fed into an AES from the US (BBC, 2009). Churchill’s Battle of Britain radio speech in 1940, known as "Their Finest Hour", did not meet with approval: its repetition of ‘upon’ and ‘our’, and its use of the word ‘might’ were picked out as problematic. (Churchill used it both as a modal verb and in the sense of ‘power’.) Golding’s *Lord of the Flies* did not follow regular sentence structure, and Hemingway showed ‘lack of care in style of writing and vocabulary’ (BBC, 2009). Shermis (2012) discussed the response of an AES to the Gettysburg Address, which received scores of two or three out of a possible six across a range of features. An initially stunned reaction turned to relief when he approached an historian who pointed out the speech is more famous for when and where it was given, than for its technical beauty, given its semi-improvised composition. (JFK’s inaugural address, carefully composed in reflective mode, rated by the same programme, gained full marks.)

AES are often, in these experiments, held to a far higher standard than human markers and their mark-schemes. If it were not for their fame, I have no doubt all of the texts mentioned in the previous paragraph would have problems identified by humans applying GCSE English mark-schemes. It should be uncontroversial that the writing that we expect our students to produce (and therefore an AES to grade) prioritises basic competencies and rule-following over the elusive (and debated) qualities of great literature. English teachers spend a great deal of time enforcing rules about writing in full sentences, while most great works of literature encompass fragments. We read Churchill's speeches through the lens of history and emotion. This changes the nature of the text for a human. It does not for a machine. If that text were produced by a student, apart from applauding his or her precocity (and anachronism!) of vocabulary and sentence structure, without those lenses, without the echo, it probably would *not* score at the very top.

Writing is a fundamentally social activity, an act of communication between two people. Deane claims that 'because all writing is social, all writing should have human readers, regardless of the purpose of writing' (2013, p. 8). Elsewhere AES have been criticised because they cannot provide 'authentic audience awareness' (Human Readers, 2013). Authentic audiences are rare in the world of educational assessment. The essay is a somewhat artificial task, authentic only in the context of school. In other English tasks the authenticity is limited because whatever the specified audience, in practice the teacher or examiner marking the task is rarely drawn from that group.

It is nevertheless true that an essay marked by an AES is not fulfilling writing's primary function of communication, because a computer, no matter how sophisticated, is not 'reading' an essay. Whether any act of examination writing deserves a human reader is considered below. But here I would like to address Deane's assertion that all writing should have a human reader, outside of the context of educational assessment. There are a number of

cases in which writing is done not for a reader other than the self: diary writing is a prime example, which may or may not be revisited by the author. In addition, there is the question of writing done for the purpose of reviewing thoughts, for sorting and considering ideas, for establishing arguments. The writing is not meant to be read; it is the act of writing which is important, not the act of communication. It is not de facto true, therefore, that all writing should have a human reader other than its author.

MARC

The intention in imagining a technically-perfect automated essay scorer, is to explore the ethical and moral issues behind AES without the trouble of technical arguments. This experiment therefore presumes an ungameable AES, with an immensely sophisticated algorithm, which is able to reward the characteristics of writing which are proxies for the deep arguments and careful construction of text. It is able to identify and check facts enough of the time to make fictional construction of facts less appealing than simply learning them. For no other reason than the pun, my automated essay scorer is named MARC. MARC is able not only to assess an essay but also to give an estimate of the likelihood of its own correctness. That is, some essays are highlighted as requiring human intervention to check that their grade is right. MARC is trained to grade any given task by the input of 100 pre-marked essays which cover the range of potential attainment.

Reliability and Validity

These technical aspects are an essential part of the morality of assessment. They are requirements for ‘fair’ examination: without reliability we cannot guarantee that the difference between any two students’ grades accurately and therefore fairly reflects the

difference in the quality of their work. Two aspects of reliability are required: intra- and inter-rater. If an examiner were to read a given student's script on the first or the twenty-first day of her examination marking, she would award it the same mark (intra-rater reliability). That student's script should also receive the same mark, within acceptable limits of variation, no matter who marks it (inter-rater reliability).

The phrase 'acceptable limits' covers a multitude of sins. The understanding of those working within the system is that marking incorporates a certain amount of subjective judgement; steps may be taken to minimise variation, but it will always exist. Public perception is there is a 'right' mark and a 'fair' system would see their child get that 'right' mark. Fundamentally, measurement of academic attainment particularly in relation to longer pieces of writing is different from measuring height. There can be no absolute right mark for an English essay; the idea of error is redundant. There can be convergence of opinion, but an essay is worth the mark an experienced examiner gives it. In England we use a Principal Examiner's judgement as the 'right' mark for aligning other people's judgements, so that the question examiners must ask themselves is 'what mark would the Principal Examiner give this script?' *not* 'what mark should this script get?' This is aligned with the idea of an AES which predicts a mark (according to analysis of previously marked scripts), rather than grading itself.

Reliability is overwhelmingly prioritised over validity; it is easier to measure and it is the aspect which is highlighted whenever a school appeals against the grades its students have received. The first major project announced by the newly formed OfQual was a review of reliability (final report Baird *et al.*, 2011): this topic dominates our discussion of assessment. If reliability is really the only aspect of examination marking we care about, then MARC wins hands down. Having formed its algorithm, it will apply it exactly the same way to every script: intra-rater reliability is absolute in a machine. It cannot have a bad day at school, or

resort to alcohol to ease the marking experience, or take offence at a candidate's stance on religion. The question of inter-rater reliability never even arises: there is only one rater.

However, it is on matters of validity that the challenge to computer-based scoring is more usually based. Validity, loosely defined, is the extent to which an assessment tests what it should be testing. The application of a test of validity to examination *marking*, therefore, might require that the judgement be formed on the basis of appropriate cues. Such validity is implied by the mark-scheme, which provides appropriate prompts for the examiner and which can be scrutinised by teachers, pupils and parents, and linked to regulator-set criteria. In the automated system, MARC will assemble its algorithm on an opaque basis, returning assessment to the black box system of the early history of examinations, before strenuous efforts to make examiners' work transparent in the second half of the 20th century. There are advantages to this. An opaque system is harder to game; there are those who would suggest that the transparency of the assessment criteria has already led to gaming of the system with a focus on 'ticking the boxes' rather than producing work of innate quality. Others would counter that this is due to misunderstanding by inexperienced teachers and examiners alike and that quality remains the abiding criterion and ticking the boxes in name but not spirit is unlikely to result in anything other than a mediocre essay.

The cues on which MARC operates are unlikely to be those of human markers. To oversimplify the case to an absurd degree, a system might include factors such as: the inclusion of eight out of ten key words; the length and syntactic accuracy of sentences; the level of vocabulary being used; and the presence of thesis words in both introductory and concluding paragraphs *inter alia*. It is almost impossible to conceive of the sophistication of the analysis of which the system is capable: the examples which are easy to conceive are the facile ones which give an inappropriate impression of what MARC would build its judgement on.

For the purposes of this though experiment, I accept the proposition that MARC's system works on proxies, and that humans do not. Is it therefore an invalid judgement? Just as human markers are actually trying to predict what grade the Principal Examiner would award, MARC is also a prediction engine first and foremost. Like a statistical regression tool, it builds a model of which factors most accurately predict the grades awarded by the Senior Examiner in the initial training period. The difference is that rather than a human building the model coherently, MARC takes the cherry-picking approach, simply selecting for itself the (massively multiple) factors which are most useful in predicting the outcomes in its training material. Providing the original training material has judgements based on valid prompts, I would argue that it does not matter that MARC is using prompts which are not necessarily related to the domain of the examination. They *are* valid factors for predicting what the Examiner would award: if her awards are valid, MARC's must also be.

Giving students their just deserts

But the question remains, *should* the essays on which students' futures depend be entrusted to MARC, even if we accept its validity and its unassailable superiority in terms of reliability? Is there a moral dimension to the procedures by which we assess secondary school students' attainment at GCSE and GCE?

Perhaps a student who has spent three hours producing an essay on WW1 literature at the end of their school career *deserves* to have that answer read and appreciated by an examiner who will then come to an expert judgement. An automated judgement does not fulfil this: no matter how advanced MARC is, it is incapable of 'appreciating' an essay and 'reading' occurs only in very specific terms.

This position contrasts MARC's system with an idealised version of examination marking by humans; the numbers of scripts which examiners are required to mark within

short periods of time mean that they are unlikely to have time to thoroughly read and 'appreciate' the essay. They are most likely to be skim-reading, and to have their attention divided at least two ways between a mark-scheme and the script, if not further by environmental distractions.

Once again we will assume that reality is close to the ideal. Does the parent or teacher who asserts the moral deserts of the pupil to have their work read carefully prioritise that over their deserving an accurate grade, produced quickly, which reflects appropriately their performance relative to that of their classmates and of the national cohort? Given an oppositional choice parents, teachers and pupils would, I think, choose the latter, particularly in terms of speed and accuracy. But it is not an oppositional choice; in the idealised version of the current system, both outcomes can apparently be satisfied.

The question of how rapidly a piece of work can be marked is an interesting one. From a learning point of view the literature is clear that rapid or instant feedback is preferred by students and is better for learning (Epstein & Brosvic, 2002; Dibattista & Gosse, 2006). Proponents of EdX's system have suggested that the real value of scoring software lies in its classroom use, allowing teachers to set essay tasks much more frequently, on the basis that writing essays is the best way to improve students' abilities. Access to this kind of assisted self-assessment, targeted at learning rather than judgement, could escape the moral imperative of having a student's work 'read'? Online practice tests, from the Driving Theory Test to Teachers QTS Skills Tests, are immensely popular with learners. The same might be true for practice grading of more complex forms of assessment, provided that it was aligned with the final human grading.

So, do students 'deserve' to have their work read – is it a moral requirement? Students produce vast quantities of work over their school lives, and there is a strong imperative that they should *not* be producing that work for nothing. On occasion, desirable or not, children

are set work in school which has no purpose other than keeping them busy, but the idea of that work going unread is complete anathema to pupils, teachers and parents. Children deserve to have their work read, and responded to. But an examination script is produced in a completely different context. It is designed not to elicit a response or stimulate feedback, but to generate certification for a particular attainment. The moral imperative seems to be for accurate judgement, produced within a reasonable time. Students *deserve* to have their effort in creating a twenty page essay rewarded appropriately; the appropriate reward is a fair grade.

What remains?

In writing this I have explored the principal logical objections to the use of an (admittedly futuristically ideal) automated essay assessment programme, and at least balanced them with the counter arguments. It is at this point that the thought experiment forces me to consider why, precisely, I do not like the idea. Logically, I am convinced, given the fulfilment of certain criteria, but there is a role for something else in human judgement, and that is affect.

First, there are some practical objections related to affect, and what exactly it is we wish to judge in examination. To take English writing at GCSE: tasks in this area are notorious for their tendency to stimulate writing that focuses on the sad, the depressing and the downright horrific. It is well known that GCSE candidates can achieve high marks in one of two ways: make the marker laugh or make the marker cry. A computer cannot laugh or cry.

It would be possible to programme the system to look for the features that students are taught to use to manipulate the feelings of the reader. But not only does reliance on such triggers lead to the possibility of ‘gaming’ the system, it also ignores the fact that successfully provoking the desired emotional reaction is a matter of balancing techniques and exercising emotional judgement. Furthermore, the extent of that success *is* a subjective experience on

the part of the reader: there is no way to make emotional reaction an objective response. This subjectivity on the part of human markers would introduce an element of instability into MARC's capacity to predict the appropriate mark. No human can predict with certainty how any other will react, stranger or closest friend, despite our empathetic capabilities; how much less can the computer who has no empathy?

Very well, then, say we exclude pieces testing the 'creative' writing of students from MARC's operations. An essay on Keats's *Nightingale* can have no affect component, surely? The sentence answers its own question: no writing is entirely divorced from emotion or opinion and this is conveyed by its construction, particularly through the rhetorical devices it employs. An essay which employs *no* rhetorical devices is also likely to be short on content and complexity.

There is a flaw in this argument. Any given human will react differently, subjectively, to the affective element of an essay, but examiners do not follow their subjective feelings in marking or should not. I can take a step back as a marker and see that a piece is written in such a way that it might make a reader laugh or cry, if they weren't marking their thirtieth essay of the day. It is precisely these elements which I can identify which MARC would also be capable of identifying, although even this best and most ideal AES would surely not be able to adjudicate the subtle levels of difference between clumsy and sophisticated application of those elements. But then the mark-schemes for GCSE English compositions do not necessarily do so either.

It is appropriate that affect should prove a sticking point for the ability of an AES to judge the quality of writing, as it is undoubtedly an affective reaction which makes individuals uneasy about the use of automated essay scoring. Zajonc suggested that human decisions were often based on affect, before they then seek 'to justify these choices by various reasons' (1980:155). The difficulty with MARC is that (for me) the justification seems

impossible, and yet the affective reaction remains. There is an element of defensiveness, perhaps, of fear that a computer will take over the marker's job. I have been an examiner, and have found that to be a useful supplement to my salary as a junior teacher; perhaps this contributes to my suspicion.

Deane (2013) has suggested that too much coverage of AES has been polarised to the extremes. He suggests a middle way, in which they have their place alongside human markers, working as a second marker, which brings the added benefits of reliability without the worries concerning validity. In the UK system dual marking has not been routinely affordable; this could therefore be a substantial improvement on the current system. It is really only in the context of large scale examinations that the costs of AES are justified, and in the context of second marking, many of the objections on principle to the use of an AES no longer hold. In the context of a classroom support to teachers, to provide faster and greater formative assessment (though without the teacher gaining the close knowledge of his or her students' work and abilities which is so essential for successful formative assessment), as part of an assessment regime rather than its whole, there are definite advantages. The full technical realisation of MARC may be some way off, but it is worth educators considering whether we wish to get to that point, what we would do with it once we did, and upon what principles our judgement is founded. We must have the debate on the basis of principles, but also need to deal in the realities of human marking rather than an idealised version of them.

Biographical Note

Victoria Elliott was previously Lecturer in English in Education at the University of York. She is now Associate Professor of English and Literacy Education at the University of Oxford. Her research covers both English and assessment, and she is an external subject

expert for Ofqual. This thought experiment stems from playing devil's advocate in discussions on this topic with students, colleagues and teachers.

References

Baird, J.-A., Beguin, A., Black, P., Pollitt, A., & Stanley, G. (2011) *The reliability programme: final report of the technical advisory group Ofqual/11/4825* (Coventry, Ofqual).

BBC (2009) *Great writers 'fail' online test* [online]. Available online at: <http://news.bbc.co.uk/1/hi/education/8356572.stm> (accessed 26th April 2013).

Deane, P. (2013) On the relationship between automated essay scoring and modern views of the writing construct, *Assessing Writing*, 18(1), 7–24.

DiBattista, D., & Gosse, L. (2006) Test anxiety and the immediate feedback assessment technique, *The Journal of Experimental Education*, 74(4), 311–328.

Elliott, V. (2011) *Marking Time: the decision-making processes of examiners of 'A' level English and History* (Unpublished DPhil thesis).

Epstein, M. L., & Brosvic, G. M. (2002) Students prefer the immediate feedback assessment technique, *Psychological reports*, 90(3c), 1136–1138.

The Guardian (2013) *Should software grade essays?* Available from <http://www.guardian.co.uk/commentisfree/poll/2013/apr/25/should-software-grade-essays-poll>. [Accessed 26th April 2013]

Human Readers (2013) *Professionals Against Machine Scoring Of Student Essays In High-Stakes Assessment* [online]. Available from <http://humanreaders.org/petition/research_findings.htm> (accessed 7th May 2013).

Larson, L. (2013) Outrage over software that automatically grades college essays to spare professors from having to assess students' work, *The Daily Mail* [online] (Last updated

07:02 on the 5th April 2013). Available at < http://www.dailymail.co.uk/news/article-2304317/Outrage-software-automatically-grades-college-essays-spare-professors-having-assess-students-work.html> (accessed 26th April 2013).

Markoff, John (2013) Essay-grading software offers professors a break, *New York Times* [online]. Available online at <http://www.nytimes.com/2013/04/05/science/new-test-for-computers-grading-essays-at-college-level.html?pagewanted=all> (accessed 4th April 2013).

Page, Ellis B. (1966) The imminence of grading essays by computers, *Phi Delta Kappan* 47, 238–243.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001) *Stumping e-rater: Challenging the validity of automated essay scoring (GRE Report, No. 98-08bP)* (Princeton: ETS) Available at <www.ets.org/Media/Research/pdf/RR-01-03-Powers.pdf.> (accessed 26th April 2013).

Sadler, D. Royce (1989) Formative assessment and the design of instructional systems, *Instructional Science*, 18, 119–144.

Shermis, M. D. (2012) *Famous speeches tested* (Youtube) Available at <http://www.youtube.com/watch?feature=player_embedded&v=1feVI1VDkZA>. (accessed 26th April 2013).

Zajonc, R. B. (1980) Feeling and thinking: Preferences need no inferences, *American Psychologist*, 35, 151–175.