Semester 1 2018

www.cpl.asn.au/journal

JPL JOURNAL OF PROFESSIONAL LEARNING

# ROBOT MARKING: AUTOMATED ESSAY SCORING AND NAPLAN – A SUMMARY REPORT

**Les Perelman critiques the accuracy of the evidence-base for computer marking and assessment in our schools…**

This summary report is written in response to proposals for employing an Automated Essay Scoring (AES) system to mark NAPLAN essays, either as the sole marker or in conjunction with separate scores from a human marker. Specifically, this summary will address assertions regarding AES's appropriateness made in *An Evaluation of Automated Scoring of NAPLAN Persuasive Writing* (ACARA NASOP Research Team, 2015) [henceforth referred to as *The Report*]. After describing the primary strategies AES systems use to compute scores of writing ability and the major studies of the efficacy of AES for high-stakes assessments, various critiques of AES are discussed. Finally, an analysis of *The Report* concludes that both its review of the literature and the study described in it are so methodologically flawed and so massively incomplete that it cannot justify any use of AES in scoring the NAPLAN essays.

## How AES works

All AES systems analyse only textual features that can be represented and manipulated mathematically (Zhang, 2013). AES, from its beginnings in the 1960's (Page, 1966) relies heavily on the use of proxies that can be easily counted. It cannot directly measure a student's adept use of vocabulary. Instead, it often just calculates the number of infrequently used words in a text (Attali & Burstein, 2006; Page, 1966). Because it cannot actually comprehend how well a topic is developed in a paragraph, it determines development by counting the number of sentences in each paragraph (Attali & Burstein, 2006; Burstein, Marcu, & Knight, 2003). And just counting the number of commas has been successfully used in helping to calculate an overall score of an essay that will match that of human readers (Bennett & Zhang, 2016; Simon, 2012).

The other methods used by AES systems consist of various natural language processing techniques. All of these techniques work by statistically identifying key words in a text and analysing their frequency, often in relation to other words. E-rater's natural language technique begins with the assumption that some of the words in high-scoring essays have a high probability of occurring in other high-scoring essays, and similarly, most low-scoring essays will contain a subset of words associated with low scores. It then employs statistical techniques based on the vocabulary in an essay to determine the essay's score category as well as the relation of the essay's vocabulary to that of the highest scoring essays (Attali & Burstein, 2006). Some techniques, such as Latent Semantic Analysis, create matrices based on single words and, like e-rater, ignore word order (Foltz, Streeter, Lochbaum, & Landauer, 2013; Landauer, Foltz, & Laham, 1998). Many AES systems, such as ETS's e-rater, use a hybrid approach that combines proxies with other machine learning and natural language processing techniques.

## Efficacy of AES

Given our current linguistic and computational knowledge, does AES work? There is already some indication that in some cases—such as writing in response to open-ended prompts, in which students have wide latitude in direction and creativity—AES cannot replicate human markers (McCurry, 2010). The most ambitious research study is the Hewlett ASAP study referenced by The Report. Although the Hewlett Study is not in any way seminal, it was extremely ambitious, using a total of 22,029 student essays based on eight different writing prompts from six U.S. state tests. These essays were divided into a Training Set, a Test Set, and a Validation Set.

The Hewlett Study Report exists in three forms: the original conference paper (Shermis & Hamner, 2012), a version that appeared in a collection of essays co-edited by the paper's first author (Shermis & Hamner, 2013), and a single-authored article that appeared in a peer-reviewed journal and that concluded with a fairly lengthy list of the study's limitations (Shermis, 2014a). [Full disclosure: I am on the editorial board of the journal.] Curiously, *The Report* references only the first two versions, ignoring the more authoritative peer-reviewed article, which is qualified in its endorsement of AES.

## *Strengths of the Hewlett Study*

One unfortunate limitation of the study was that the agreement with the vendors prohibited the research group from conducting any statistical tests comparing the vendor and human marker scores (Bennett & Zhang, 2016; Rivard, 2013). However, the study report (in all three versions) was thorough in presenting demographic statistics for each of the U.S. states participating in the study as well as statistics in two general categories:

- **Descriptive statistics** such as the number (N), mean, and standard deviation (STD) on each essay set for human markers and all nine vendors.

- **Measures of agreement** such as percentage of exact agreement, percentage of exact plus adjacent agreement, Cohen's kappa, Quadratic-weighted kappa, and the Pearson product-moment correlation coefficient.

The research team also subsequently released the raw scores on the Test Set for seven of the nine vendors for confirmation and analysis. Two vendors did not want their data made public even though the sets were anonymous.

## *Limitations and critiques of the Hewlett Study*

The Hewlett Study results were released with much fanfare. The University of Akron reported

> A direct comparison between human graders and software designed to score student essays achieved virtually identical levels of accuracy, with the software in some cases proving to be more reliable, a groundbreaking study has found. ("Man and machine: Better writers, better grades," 2012)

Yet close analysis of the data casts doubt on that claim as well as raises questions about major methodological elements of the study:

- The data do not support the claim that machines were able to match human readers. Indeed, analyses of the specific data tables indicate that humans possessed higher levels of accuracy than machines (Bennett, 2015; Bennett & Zhang, 2016; Perelman, 2013, 2014). The exhaustive analysis of Bennett (ETS's Norman O. Frederiksen Chair in Assessment Innovation) and

Zhang (2016), in particular, refutes any claim that the AES scores in the Hewlett Study matched the reliability of human readers.

- Five of the eight data sets consisted of paragraphs not essays, with mean lengths of 99–173 words (Shermis, 2014a; Shermis & Hamner, 2012, 2013).

- The four essay sets in which the machines performed best (Sets 3, 4, 5, and 6)
  - were not marked on writing ability but solely on content;
  - had reliability assessed using the higher of the two human markers' scores, producing different scoring formulas for machines and humans, which made any comparison problematic and privileged machines (Bennett, 2015; Bennett & Zhang, 2016; Perelman, 2013, 2014). The importance of this last assertion, however, has been contested (Shermis, 2014b).

- Only two of the eight essay sets in the study employed, like NAPLAN, a composite score based on a combination of analytic scores. The machines performed poorly in comparison to humans for these sets (Shermis, 2014a; Shermis & Hamner, 2012, 2013)

## Critiques of AES

One major failing of *The Report* is that it completely ignores the significant body of scholarship critical of various applications of AES. The focus here will be on those objections that are the most relevant to NAPLAN. For a more complete listing of some excellent collections of essays on AES see *Appendix A*.

## Lack of rhetorical situation

One of the most common objections is that writing is communication, the transfer of thoughts from one mind to another. As various scholars have noted, AES creates a non-rhetorical situation (Anson, 2006; Condon, 2006, 2013; Ericsson, 2006; Herrington & Moran, 2001, 2012). Students are writing not to inform, entertain, or persuade another mind; they are writing to an entity that can only count. In essence, the audience has been replaced by a machine. Even in cases in which there is both a human and a machine marking the essay, the student will be aware that half the score is coming from an entity that does not understand meaning but is simply looking for specific elements. Students then have a dual audience; they must produce a text that will satisfy the machine, even if a human reader is also present.

## Reductive

Because AES is solely mathematical, it cannot assess the most important elements of a text. The following paragraph is not written by critics of AES but by its developers, including three very senior individuals at the Educational Testing Service and four vice presidents at Pearson Education and Pearson Knowledge Technologies:

> Automated essay scoring systems do not measure all of the dimensions considered important in academic instruction. Most automated scoring components target aspects of grammar, usage, mechanics,

> spelling, and vocabulary. Therefore, they are generally well-positioned to score essays that are intended to measure text-production skills. Many current systems also evaluate the semantic content of essays, their relevance to the prompt, and aspects of organization and flow. Assessment of creativity, poetry, irony, or other more artistic uses of writing is beyond such systems. They also are not good at assessing rhetorical voice, the logic of an argument, the extent to which particular concepts are accurately described, or whether specific ideas presented in the essay are well founded. Some of these limitations arise from the fact that human scoring of complex processes like essay writing depend, in part, on "holistic" judgments involving multivariate and highly interacting factors. This is reflected in the common use of holistic judgments in human essay scoring, where they may be more reliable than combinations of analytic scores. (Williamson et al., 2010 p. 2)

This passage makes two points extremely relevant to the use of AES in marking NAPLAN. First, AES cannot assess some of the key criteria addressed by the NAPLAN writing test, such as audience, ideas, and persuasive devices (i.e. the logic of an argument). Second, AES is more reliable providing a single holistic score rather than the sum of analytic scores, such as the ten trait scores of the NAPLAN. This second point is supported by how the essay portions of two high-stakes American tests, the new SAT Essay and the Analytical Writing Essays of the Graduate Record Examination (GRE), are marked. The new SAT Essay is marked on three analytic categories, which are not combined but reported separately. The analytic scores are produced by two human markers (College Board, 2017). The GRE Essays, on the other hand, are evaluated by a single holistic score for each essay and are marked both by a machine and by a human (Educational Testing Service, 2017).

## *Weaknesses in grammatical analysis*

The above passage from AES developers, like similar claims (Deane, 2013), assumes that AES systems are precise in identifying grammatical errors. However, anyone who has ever used a grammar checker suspects that this is not the case. English grammar, like the grammar of any natural human language, is extremely complex and interdependent on such factors as meaning and context. AES grammar checkers miss many grammatical errors (False Negatives), while classifying perfectly grammatical constructions as errors (False Positives). When analyzing 5,000 words of an essay by Noam Chomsky originally published in *The New York Review of Books*, the grammar checker modules of ETS's e-rater identified 62 grammatical or usage errors, including 15 article errors and 5 preposition errors (Perelman, 2016). None of them were actually errors.[1] In addition, AES grammar checkers often focus on grammatical non-problems, such as beginning a sentence with a coordinating conjunction, possibly because such constructions are very easy for a machine to identify.

One of the most complex linguistic features of English is the set of rules governing the use of articles; these rules are especially challenging for speakers of languages such as Mandarin or Russian that do not have articles. Computational linguistic models of English article use are disappointing. One model, for example, deployed in 2005, could detect 80% of article errors with a False Positive rate of approximately 50% or detect only 40% of article errors but reduce the False Positives to 10% (Han, Chodorow, & Leacock, 2006). A comparison of error identification by two instructors and e-rater 2.0 of 42 English Language Learners' papers demonstrated that e-rater is extremely inaccurate in

[1] All of the examples are from ETS's e-rater simply because other vendors no longer allow academic researchers access. A Pearson vice president responded to a reporter's request to allow me access to the Intelligent Essay Assessor by refusing and stating, "He wants to show why it doesn't work" (Winerip, 2012).

identifying the types of major errors made by ELL, bilingual, and bidialectical students. The instructors coded 118 instances of missing or extra articles; e-rater marked 76 instances, but 31 of those (40.8%) were either False Positives or misidentified (Dikli & Bleyle, 2014). The current inability to develop reliable grammar checkers is best exemplified by the decision of Microsoft Research, one of the largest software companies in the world, to discontinue its ESL Assistant Project (Gamon, 2011). AES is inaccurate and unreliable at assessing even low-level writing traits such as grammatical correctness.

## *Fairness*

Related to grammar is the issue of fairness. Do AES machines treat all linguistic, national, and ethnic groups the same? Two reports by the Educational Testing Service (Bridgeman, Trapani, & Attali, 2012; Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012) indicate that in the essay portions of both the Test of English as a Foreign Language and the GRE, the e-rater scoring engine gave significantly higher marks to native Mandarin speakers, especially those from mainland China, than did human markers. In some instances, the difference between the machine score and human was very large, close to 0.40 of a standard deviation. Conversely, in some instances, African-Americans, particularly males, were given significantly lower marks by e-rater than they were by human markers. Another study reported that Vantage Technology's ACCUPLACER, which has an essay section scored by the IntelliMetric scoring engine, underpredicted portfolio and final course grades for African-American and Hispanic students (Elliot, Deess, Rudniy, & Joshi, 2012).

Possibly, the unevenness of the grammatical components of the scoring engines contributes to the machines' under- and overreporting marks. Native Mandarin speakers and native speakers of other languages that do not have articles make more errors in the use of English articles than speakers of languages that employ articles. Because grammar detectors perform so poorly in correctly identifying English article usage, they may be contributing to the machines' inflating the scores of Mandarin speakers. One prominent feature of African-American dialects of English is a difference in verb constructions. These constructions are easy for a machine to identify and may be overcounted in comparison to the response of a human marker. Another possible explanation is that people from mainland China receive extensive coaching for these tests and may be including memorized passages that appear more relevant to a machine than they do to a human marker (Bridgeman et al., 2012).

Whatever the explanation, unfairness by machines in inflating the marks of some linguistic groups and artificially lowering the marks of others is morally indefensible and, possibly, illegal. Before any AES system is deployed, extensive research is needed to ensure that the machines do not penalize or privilege specific linguistic communities.

## *Construct-Irrelevant Response Strategies (Gaming)*

Because AES relies so heavily on proxies in marking, various studies have shown that AES machines are extremely vulnerable to construct-irrelevant response strategies, that is, providing the machine with the proxies it employs without actually displaying the traits of good writing that they are supposed to represent.

For most AES machines, the strongest single proxy is length (Perelman, 2012, 2014). As noted previously in the discussion on fairness, it appears that tutors in mainland China have students memorize sentences that they then insert in essays to increase their score (Bridgeman et al., 2012). Although ETS is attempting to develop tools to catch such gaming strategies (Bejar, Vanwinkle, Madnani, Lewis, & Steier, 2013), they appear still to be effective (Bejar, Flor, Futagi, & Ramineni, 2014; Powers, Burstein, Chodorow, Fowles, & Kukich, 2001).

Perhaps the most theatrical example of the vulnerability of AES systems to gaming strategies is the BABEL Generator developed by the author and three undergraduates from Harvard and the Massachusetts Institute of Technology (Kolowich, 2014). Just by randomly creating nonsense sentences with long, rarely used words and occasionally peppered with synonyms of at most three topic words, the BABEL Generator is able to create essays that receive high scores from AES machines such as e-rater and Vantage Technology's IntelliMetric. Two pairs of top scoring, BABEL-written GRE essays along with a link to the BABEL Generator are displayed in Appendix B.

The main danger, however, is not from absurd machines such as the Babel Generator, but from the implications of such stumping studies. That which is tested will be taught. If wordy essays with long sentences and obscure vocabulary will produce high scores on high-stakes tests, that is what teachers will be emphasizing. Rather than improve the writing ability of students, AES may well encourage the production of verbose, high-scoring gibberish.

## Inaccuracies, methodological flaws, incomplete information, and anomalies in an evaluation of automated scoring of NAPLAN persuasive writing

The flaws in *The Report* and the study it describes are so major that it cannot justify any use of AES in high-stakes testing situations.

### *Inaccuracies*

The most egregious mistake in *The Report* is in the account of the Hewlett competition on page 5: "The rate of agreement was higher between any of the automated scoring engines and human markers than that between the two human markers." Even a cursory examination of the data in any of the three papers reporting on the study reveals the gross inaccuracy of this statement (Shermis, 2014; Shermis & Hamner, 2013). As Bennett and Zhang (2016) demonstrated, humans actually performed more reliably. The most vivid refutation of this claim can be made by comparing the human–human reliability to the human (resolved score)–machine reliability for each of the metrics for each of the essay sets and for just one scoring engine, MetaMetrics's Lexile Writing Analyser. Table 1 displays this comparison. Rather than being more reliable than the human markers, Lexile is substantially less reliable for every metric and essay set except for two of the metrics for Essay Set 8 (shaded). Lexile was chosen for several reasons. First, its performance was the poorest of any of the scoring engines. Second, it is one of the four engines used in the study described in *The Report*. Finally, unlike the other engines, Lexile is not trained for a specific prompt but, instead, measures a general trait, text complexity (*The Report*, p. 7)

**Table 1: Comparison of Agreement Metrics Between the Two Human Markers (H-H) and Between MetaMetrics's Lexile Writing Analyser and Human Markers**

| Essay Sets | Exact Agreement | | Kappa | | Quadratic-Weighted Kappa | | Correlation Pearson *r* | |
|---|---|---|---|---|---|---|---|---|
| | H1 - H2 | Lexile | H1 – H2 | Lexile | H1 – H2 | Lexile | H1 – H2 | Lexile |
| 1 | 0.64 | 0.31 | 0.45 | 0.16 | 0.73 | 0.66 | 0.73 | 0.66 |
| 2A | 0.76 | 0.55 | 0.62 | 0.30 | 0.80 | 0.62 | 0.80 | 0.62 |
| 2B | 0.73 | 0.55 | 0.56 | 0.27 | 0.76 | 0.55 | 0.76 | 0.55 |
| 3 | 0.72 | 0.63 | 0.57 | 0.45 | 0.77 | 0.65 | 0.77 | 0.65 |
| 4 | 0.76 | 0.47 | 0.65 | 0.30 | 0.85 | 0.67 | 0.85 | 0.68 |
| 5 | 0.59 | 0.47 | 0.44 | 0.28 | 0.74 | 0.64 | 0.75 | 0.65 |
| 6 | 0.63 | 0.51 | 0.45 | 0.31 | 0.74 | 0.65 | 0.74 | 0.66 |
| 7 | 0.28 | 0.07 | 0.18 | 0.03 | 0.72 | 0.58 | 0.72 | 0.58 |
| 8 | 0.29 | 0.08 | 0.16 | 0.04 | 0.61 | 0.63 | 0.61 | 0.62 |

Source: Shermis, 2014a, Tables 7, 9, 10, and 11

Another major problem is the citation of Attali (2013). Attali does indeed offer practical advice on validity in writing assessment. The advice he offers, however, is contrary to the conclusions of *The Report*. He argues that AES is severely limited and cannot assess several of the NAPLAN traits. He states,

> we believe that a serious consideration of the construct argument against AES should lead one to accept its basic premise—because the machine is not able to read the essay, it will not be able to assess such aspects as the quality of argumentation or the development of characters in a narrative, as human readers do. . . . We believe that AES should be based on an alternative definition of its intended use. Specifically, it should be constructed primarily as a *complement* to (instead of a replacement for) human scoring, *limited* in its ability to measure a subset of the writing construct. (p. 182)

*The Report* also contains problems in terminology. Attali employs the term *construct* correctly. At its conclusion, however, *The Report* defines *construct validity* in this passage "ACARA will examine if the introduction of automated scoring has an effect on the substance and quality of student writing ('construct validity')" (p. 14). Construct validity is a complex and evolving concept. At its core, however, is the key concept that the measure is representing the abstract ability (the construct) that it is claiming to assess. Thus "the substance and quality of student writing" is the construct. The question is whether AES can faithfully measure it, not whether AES can affect it.

Another problem with terminology is the misuse of the term *lexical.* The term is correctly defined in footnote 2 on page 4. On the following page, however, the "lexical properties of essays" are listed as "sentence structure, paragraphing, punctuation and spelling." These elements of writing have little or anything to do with the term *lexical.*

A final problem with language is the use of the term *cognitive interview.* Since this term in all Anglophone countries usually refers to a specific technique used in forensic investigations (Davis, McMahon, & Greenwood, 2005), it is extremely unclear what *cognitive interview* means in this context.

## *Methodological flaws and incomplete information*

While the inaccuracies in the report were disconcerting, it is the study's very flawed methodology accompanied by a consistent lack of definition and detail that make *The Report* inappropriate in justifying any decision to employ AES in marking the NAPLAN.

### A Convenient Sample Defines a Pilot Study

The method section of *The Report* states "A single persuasive prompt was administered to a convenient sample of year 3, 5, 7 and 9 students as part of a larger online assessment study" (p. 6). Major studies, especially those with national consequences usually employ a *representative* sample, or, if it is a large, broadly-based sample, possibly a *random* sample. In research, convenience sampling is limited to pilot studies because of the risk of sampling errors. The Discussion section of *The Report* makes it clear that this study is a pilot and that there will be larger follow-up studies: "ACARA will next expand its research to include larger samples of students and multiple prompts within and across writing genres that NAPLAN assesses (persuasive and narrative)" (p. 13). The plan for future research is also explicitly stated in the August 13, 2016 ACARA web page on Automated Essay Scoring:

> More research is planned for 2016 which will include a larger sample of students, multiple prompts within and across writing genres (persuasive and narrative) and key validity questions—does the use of AES affect features of student writing and writing instruction—to inform a recommendation to Education Ministers about the approach to be used in 2017.

The current version of the web page omits any reference to larger follow-up studies. There is no explanation of why ACARA never undertook these crucial additional projects.

Both versions, however, claim that the sample was "broad," although there is no attempt to show that the sample was representative of the national population. Indeed, the Test Set consisted of 339 essays.

If they were evenly divided among Years, they would consist of only 110 essays for three Years and 109 for one Year.

The Method section reports the mean essay length and median raw scores by Year. These numbers appear to be for all three sets—Training, Validation, and Test Sets—although that is not certain. There is no explanation why the mean is given for essay length and the median for raw score. There also needs to be much more supporting data. The means of the Test Set for essay length and for each trait score should have been provided, along with the standard deviations for each. These numbers then needed to be compared with national statistics to ensure that the sample was representative.

Moreover, with such a small sample size, it is impossible to determine if any of the AES machines gave higher or lower scores to members of specific linguistic or ethnic groups than the scores given by human markers. Finally, there has been no evaluation of machines evaluating narrative essays.

Even more troubling is that this pilot was based on a testing format different from that currently used for the NAPLAN essay. There are now separate prompts for Years 3 & 5 and for Years 7 & 9. There has been no attempt to assess how well the machines perform on the different prompts for these two groups. The mean statistics for essay length alone indicates that length alone clearly differentiates them. Will separating these two groups make scoring more difficult for machines? This crucial question remains unanswered.

One very bizarre aspect of the study's methodology is allowing each vendor to report its results differently. The Hewlett Study, which is referred to as "seminal," correctly reported all vendor data homogeneously. Why were vendors in this study allowed to choose how they would present their data? Why are all the presentations different? Was there a deliberate attempt to avoid comparisons?

There is also some uncertainty about exactly when the vendors received the marks of the human scorers for the Test Set. On page 7, *The Report* first states that "Contractors were not provided with any marking data for these essays." At the bottom of the same page, however, it states, "Vendors completed the scoring and provided ACARA with a research report outlining the methods used in their investigation and its key outcomes." Vendors needed the Test Set marking data before they wrote a research report that included outcomes. Did they first provide ACARA with a dataset of their scores before they received the human scores? If so, this fact should have been stated explicitly.

There is also too much reliance on undefined and vague hearsay evidence. At the beginning, *The Report* states,

> Markers who scored the essays observed that student responses were at least as long, on average and of comparable quality, as those produced in paper-based tests. Even at Year 3, student lack of typing ability was not found to be a barrier to completing the task. (p. 3)

Comparing word counts of the sample to national word counts for each Year would have provided a much more accurate assessment of the effect of a computer-based test on text production. Similarly, a statistical comparison of total scores and trait scores could verify the markers' impressions with hard data. There is the statement, "psychometric analyses confirmed that the underlying writing scales performed in a similar manner to their paper-based analogues." However, that is the only reference to those analyses, which, along with supporting data, should have been an integral and substantial part of the document.

*The Report* also states,

> Invigilator observations and follow-up discussions ("cognitive interviews") with students confirmed that students were able to complete the writing task within the allotted time, without being unduly constrained by level of keyboarding skill. (p. 3)

Although probably not "cognitive interviews," it is clear that interviews did take place. What were the exact questions asked? Was there an interview protocol? In addition, it is difficult to believe that all students reported that they were not "unduly constrained." Were there some complaints? If so, how many? What was their nature?

### *Anomalies*

As mentioned, all four vendors were part of the Hewlett Competition. As also stated previously, the Lexile Writing Analyser was the poorest performer in the Hewlett Competition and it employs a generic algorithm that does not consider the specific prompt or topic. In many of the metrics, its performance was especially dismal for Essay Sets 7 and 8, the only sets in the Hewlett Competition that, like NAPLAN, employ analytical scales. Yet the Quadratic-weighted kappas in Table 3 of *The Report* indicate that Lexile performed extremely well in its ratings for Audience and Ideas, even though it did not know or consider the specific writing task, prompt, or question being posed. The current web site states "Of especial significance, the AES systems were even able to match human markers on the 'creative' rubric criteria: *audience* and *ideas*." That the machine was able to evaluate the quality of an answer to a question without knowing the question is indeed of special significance.

Moreover, although the labelling in Table 3 is unclear (and appears to include references to an Excel spreadsheet [Columns AM through AX; Columns C through Y]), it seems that the Quadratic-weighted kappa comparing Lexile results with the human marks is either 0.8828 or 0.9190. However, neither number matches those of the AES machines displayed in Table 5. There may be an explanation for these differences, but if it exists, it needs to be made explicit.

## Conclusion

Even some of the strongest proponents and developers of AES have conceded that it cannot assess high-level traits such as quality and clarity of ideas. These traits comprise the focus and reason for human communication. They need to be assessed and assessed well. The pilot study described in *The Report*, with its large amounts of hearsay evidence, extremely dubious methodology, and incorrect information, cannot justify any sort of national implementation. Before any kind of AES system is deployed either as a sole marker or in dual markings with humans, a number of issues need to be addressed:

- Evidence needs to be provided that the correct constructs are being measured by the machines. As Mark D. Shermis (2014a), the principal investigator of the Hewlett Competition, writes in the final, peer-reviewed version of his study,

  > A predictive model may do a good job of matching human scoring behaviour, but for reasons unrelated (or unsatisfactorily related) to the construct of interest. If accurate predictions of score are achieved by features and methods that do not bear any plausible relationship to the competencies and construct that the item aims to assess, then this prediction, accurate as it may be, is not sufficiently representative of the construct to warrant test use. (p. 74)

  In particular, given the relative recent unanimity among AES developers and critics of AES that the machines are incapable of reliably assessing high-level constructs, substantial evidence must be provided that machines are capable of evaluating such constructs. Without such proof, machine scoring may produce situations in which teachers, to protect themselves and their schools, spend significant time teaching students strategies to "game" the machines with construct-irrelevant strategies that will improve their scores but make their writing less effective. Possibly, independent investigators should be allowed to test the construct relevance of the machine through various types of Reverse Turing Tests and Stumping Studies.

- Given the research findings in the United States that at least one AES machine appears to overscore one linguistic group and underscore another, no AES system should be deployed until extensive pilot testing has demonstrated that AES does not discriminate against any linguistic group or groups.

- ACARA needs to provide substantial evidence, more than the poorly designed and executed pilot study, to demonstrate that AES, which has been developed primarily to generate holistic scores, can reliably score ten analytic traits.

- *The Report* and the original language on the ACARA web site stated that more extensive studies would be conducted, including ones involving the marking of narrative prompts. Given that a narrative prompt has recently been used on the NAPLAN, it is imperative that ACARA conduct studies to demonstrate that the AES systems are capable of effectively scoring the ten trait categories of the NAPLAN narrative essay.

- As noted above, the NAPLAN has changed significantly since the 2012 sample used in the pilot. There are now separate prompts (and probably separate scoring) for Years 3 & 5 and for Years 7 & 9. This change creates an entirely different scoring situation. ACARA needs to conduct pilots demonstrating that the AES machines are capable of accurately scoring these two separate groups with two different prompts.

- ACARA needs to assess the technical and keyboard capabilities of all students, including Third Year students and students from disadvantaged backgrounds, before deploying an online essay test. If text production among these groups is hindered by lack of keyboarding or technical skills, online assessment should not be deployed.

- Finally, there should be considerably more transparency and independence in these necessary research studies than was demonstrated in *The Report*. Preferably, independent investigators should constitute part of the research team.

Until these critical studies are completed and carefully evaluated, it would be extremely foolish and possibly damaging to student learning to institute machine grading of the NAPLAN essay, including dual grading by a machine and a human marker.

**Works Cited:**

*ACARA NASOP Research Team. (2015). An evaluation of automated scoring of NAPLAN persuasive writing. Retrieved from http://nap.edu.au/_resources/20151130_ACARA_research_paper_on_online_automated_scoring.pdf*

*Anson, C. (2006). Can't touch this: Reflections on the servitude of computers as readers. In P. Ericsson & R. H. Haswell (Eds.), Machine scoring of human essays (pp. 38–56). Logan, UT: Utah State University Press. Retrieved from http://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1138&context=usupress_pubs*

*Attali, Y. (2013). Validity and reliability in automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), Handbook of automated essay evaluation: Current applications and new directions (pp. 181–198). New York, NY: Routledge.*

*Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. The Journal of Technology, Learning and Assessment, 4(3).*

*Bejar, I. I., Flor, M., Futagi, Y., & Ramineni, C. (2014). On the vulnerability of automated scoring to construct-irrelevant response strategies (CIRS): An illustration. Assessing Writing, 22, 48–59. Retrieved from http://www.sciencedirect.com/science/article/pii/S1075293514000257*

*Bejar, I. I., Vanwinkle, W., Madnani, N., Lewis, W., & Steier, M. (2013). Length of textual response as a construct-irrelevant response strategy: The case of shell language. Princeton NJ. Retrieved from http://origin-www.ets.org/Media/Research/pdf/RR-13-07.pdf*

*Bennett, R. E. (2015). The changing nature of educational assessment. Review of Research in Education, 39(1), 370–407. doi:10.3102/0091732X14554179*

*Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In Technology in testing: Improving educational and psychological measurement (pp. 142–173). Washington, DC: National Council on Measurement in Education.*

*Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. Applied Measurement in Education, 25(1), 27–40. doi:10.1080/08957347.2012.635502*

*Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. IEEE Intelligent Systems, 18(1), 32–39. Retrieved from http://people.cs.pitt.edu/~huynv/research/argument-mining/Finding%20the%20WRITE%20stuff%20Automatic%20identification%20of%20discourse%20structure%20in%20student%20essays.pdf*

*College Board. (2017). SAT essay scoring. Retrieved September 15, 2017, from https://collegereadiness.collegeboard.org/sat/scores/understanding-scores/essay*

*Condon, W. (2006). Why less is not more: What we lose by letting a computer score writing samples. In P. F. Ericsson & R. H. Haswell (Eds.), Machine scoring of human essays: Truth or consequences (pp. 211–220). Logan, UT: Utah*

State University Press. Retrieved from http://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1138&context=usupress_pubs

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? Assessing Writing, 18(1), 100–108. Retrieved from http://www.sciencedirect.com/science/article/pii/S1075293512000505

Davis, M. R., McMahon, M., & Greenwood, K. M. (2005). The efficacy of mnemonic components of the cognitive interview: Towards a shortened variant for time-critical investigations. Applied Cognitive Psychology, 19(1), 75–93. Retrieved from https://www.researchgate.net/profile/Marilyn_Mcmahon/publication/216569762_The_efficacy_of_mnemonic_components_of_the_cognitive_interview_Towards_a_shortened_variant_for_time-critical_investigations/links/0046353a0f3494eed3000000/The-efficacy-of-mnemonic-components-of-the-cognitive-interview-Towards-a-shortened-variant-for-time-critical-investigations.pdf

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. Assessing Writing, 18(1), 7–24. Retrieved from http://www.sciencedirect.com/science/article/pii/S1075293512000451

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? Assessing Writing, 22, 1–17. Retrieved from http://www.sciencedirect.com/science/article/pii/S1075293514000221

Educational Testing Service. (2017). How the GRE tests are scored. Retrieved September 15, 2017, from https://www.ets.org/gre/institutions/scores/how/

Elliot, N., Deess, P., Rudniy, A., & Joshi, K. (2012). Placement of students into first-year writing courses. Research in the Teaching of English, 46(3). 285-313. Retrieved from http://www.ncte.org/journals/rte/issues/v46-3

Ericsson, P. F. (2006). The meaning of meaning: Is a paragraph more than an equation? In P. F. Ericcson & R. H. Hasswell (Eds.), Machine scoring of human essays: Truth or Consequences (pp. 28–37). Logan, UT: Utah State University Press. Retrieved from http://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1138&context=usupress_pubs

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), Handbook of automated essay evaluation: Current applications and new directions (pp. 68–88). New York, NY: Routledge.

Gamon, M. (2011). ESL Assistant discontinued. Retrieved September 20, 2017, from https://blogs.msdn.microsoft.com/eslassistant/

Han, N.-R., Chodorow, M., & Leacock, C. (2006). Detecting errors in English article usage by non-native speakers. Natural Language Engineering, 12(2), 115. Retrieved from http://ai2-s2-pdfs.s3.amazonaws.com/ee45/5537e727cf921263540134949b7042ba6521.pdf

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? College English. Retrieved from http://www.jstor.org/stable/378891

Herrington, A., & Moran, C. (2012). Writing to a machine is not writing at all. In N. Elliot & L. Perelman (Eds.), Writing assessment in the 21st century: Essays in honor of Edward M. White (pp. 219–232). New York, NY: Hampton Press.

Kolowich, S. (2014, April 28). Writing instructor, skeptical of automated grading, pits machine vs. machine. The Chronicle of Higher Education. Retrieved from http://www.chronicle.com/article/Writing-Instructor-Skeptical/146211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. Discourse Processes, 25, 259–284. Retrieved from http://lsa.colorado.edu/papers/dp1.LSAintro.pdf

Man and machine: Better writers, better grades. (2012, April 12). University of Akron News. Akron, OH. Retrieved from http://www.uakron.edu/im/online-newsroom/news_details.dot?newsId=40920394-9e62-415d-b038-15fe2e72a677

*McCurry, D. (2010). Can machine scoring deal with broad and open writing tests as well as human readers? Assessing Writing, 15(2), 118–129. Retrieved from http://www.sciencedirect.com/science/article/pii/S1075293510000218*

*Page, E. B. (1966). The imminence of grading essays by computer. The Phi Delta Kappan. Phi Delta Kappa International. Retrieved from http://www.jstor.org/stable/20371545*

*Perelman, L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). In A. Bazerman, C; Dean, C; Early, J; Lunsford, K; Null, S; Rogers, P; Stansell (Ed.), International advances in writing research (pp. 121–131). Fort Collins, CO: The WAC Clearinghouse and Parlor Press. Retrieved from https://wac.colostate.edu/books/wrab2011/chapter7.pdf*

*Perelman, L. (2013). Critique of Mark D. Shermis & Ben Hamner: "Contrasting state-of-the-art automated scoring of essays: Analysis." The Journal of Writing Assessment, 6(1). Retrieved from http://journalofwritingassessment.org/article.php?article=69*

*Perelman, L. (2014). When "the state of the art" is counting words. Assessing Writing, 21, 104–111. Retrieved from http://www.sciencedirect.com/science/article/pii/S1075293514000233*

*Perelman, L. (2016). Grammar checkers do not work. WLN: A Journal of Writing Center Scholarship, 40(7–8), 11–20. Retrieved from http://lesperelman.com/wp-content/uploads/2016/05/Perelman-Grammar-Checkers-Do-Not-Work.pdf*

*Powers, D. E., Burstein, J., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). Stumping e-rater: Challenging the validity of automated essay scoring. Princeton NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RR-01-03-Powers.pdf*

*Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). Evaluation of the e-rater® scoring engine for the GRE® issue and argument prompts ETS RR--12-02. Retrieved from https://www.ets.org/Media/Research/pdf/RR-12-02.pdf*

*Rivard, R. (2013, March 15). Humans fight over robo-readers. Inside Higher Education. Retrieved from https://www.insidehighered.com/news/2013/03/15/professors-odds-machine-graded-essays*

*Shermis, M. D. (2014a). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. Assessing Writing, 20, 53–76. Retrieved from https://assets.documentcloud.org/documents/1094637/shermis-aw-final.pdf*

*Shermis, M. D. (2014b). The challenges of emulating human behavior in writing assessment. Assessing Writing, 22, 91–99. Retrieved from http://www.sciencedirect.com/science/article/pii/S1075293514000373*

*Shermis, M. D., & Hamner, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. Retrieved August 28, 2017, from https://web.archive.org/web/20150810190434/www.scoreright.org/NCME_2012_Paper3_29_12.pdf*

*Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), Handbook of automated essay evaluation: Current applications and new directions (pp. 313–353). New York: Routledge.*

*Simon, S. (2012, March 2). Robo-readers: The new teachers' helper in the U.S. Reuters. Retrieved from http://www.reuters.com/article/us-usa-schools-grading-idUSBRE82S0ZN20120329*

*Williamson, D. M., Bennett, R. E., Lazer, S., Bernstein, J., Foltz, P. W., Landauer, T. K., … Way, W. D. (2010). Automated scoring for the assessment of Common Core standards. Retrieved from https://www.ets.org/s/commonassessments/pdf/AutomatedScoringAssessCommonCoreStandards.pdf*

*Winerip, M. (2012, April 23). Facing a robo-grader? No worries. Just keep obfuscating mellifluously. New York Times. Retrieved from http://www.nytimes.com/2012/04/23/education/robo-readers-used-to-grade-test-essays.html*

*Zhang, M. (2013). Contrasting automated and human scoring of essays (R&D Connections No. 21). Princeton NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RD_Connections_21.pdf*

*Les Perelman is an internationally recognized expert in writing assessment and the application of technologies to assess writing. He has written opinion pieces for The Boston Globe, The Washington Post, and The Los Angeles Times. He has been quoted in The New York Times, The New Yorker, The Chicago Tribune, The Boston Globe, The Los Angeles Times, and other newspapers. Dr. Perelman has been interviewed on television by ABC, MSNBC, and NHK Japan Public Television and interviewed on radio by National Public Radio, various NPR local stations, the Canadian Broadcasting Corporation, and the Australian Broadcasting Corporation.*

*The President of the College Board has credited Dr. Perelman's research as a major factor in his decision to remove and replace the Writing Section of the SAT. Dr. Perelman is a well-known critic of Automated Essay Scoring. To demonstrate the inability of Robo-graders to differentiate writing from gibberish, he and three undergraduates developed the BABEL Generator, which produces verbose and pretentious nonsense that consistently receives high marks from AES machines.*

*Dr. Perelman received his B.A. in English Language and Literature from the University of California, Berkeley, and his M.A. and Ph.D. in English from the University of Massachusetts. After a three-year postdoctoral fellowship in Rhetoric and Linguistics at the University of Southern California, Dr. Perelman moved to Tulane University where he served as an Assistant Professor of Rhetoric, Linguistics, and Writing; Director of First-Year Writing; Director of the Writing Center; and a Member of the Graduate Faculty.*

*For the next twenty-five years Dr. Perelman was Director of Writing Across the Curriculum in Comparative Media Studies/Writing at the Massachusetts Institute of Technology and served as an Associate Dean in the Office of the Dean of Undergraduate Education. He was Project Director and co-Principal Investigator for a grant to MIT from the National Science Foundation to develop a model Communication-Intensive Undergraduate Program in Science and Engineering. He served as Principal Investigator for the development of the iMOAT Online Assessment Tool funded by the MIT/Microsoft iCampus Alliance.*

*Dr. Perelman has served as a member of the Executive Committee of the Conference on College Composition and Communication, the post-secondary organization of the National Council of Teachers of English, and co-chaired the Committee on the Assessment of Writing. He is currently a member of the editorial board of Assessing Writing.*

*Dr. Perelman has been a consultant to over twenty colleges and universities on the assessment of writing, program evaluation, and writing-across-the-curriculum. Dr. Perelman has served as a consultant for writing program assessment and development for the Fund for the Improvement of Postsecondary Education of the U.S. Department of Education and for the Modern Language Association. In 2012–2013, he served as a consultant to Harvard College and as co-principal investigator in a major two-year study assessing the writing abilities of undergraduates at the college.*

*Dr. Perelman co-edited the volume Writing Assessment in the 21st Century and he is the primary author of the first web-based technical writing handbook, The Mayfield Handbook of*

***Technical and Scientific Writing.* He has published articles on writing assessment, technical communication, computers and writing, the history of rhetoric, sociolinguistic theory, and medieval literature, and he co-edited The Middle English Letter of Alexander to Aristotle.**