AN EVALUATION OF AUTOMATED SCORING OF NAPLAN PERSUASIVE WRITING

ACARA NASOP Research Team

2015

acara AUSTRALIAN CURRICULUM, ASSESSMENT AND REPORTING AUTHORITY

AN EVALUATION OF AUTOMATED SCORING OF NAPLAN PERSUASIVE WRITING

ACARA NASOP Research Team

30 November 2015

The National Assessment and Surveys Online Program (NASOP), funded by the Australian Government, was designed to support the delivery of the National Assessment Program (NAPLAN and NAP Sample) online. ACARA developed a comprehensive research program to address a range of transition issues, including test design and impacts on student performance, measurement and reporting. The research program consisted of a number of discrete but interdependent studies: a pilot study (2012), a trialling study (2013) and a development study (2014).[1]

As part of the 2012 pilot, a sample of Australian students completed a NAPLAN persuasive writing task using a computer to write their essay. The writing task was the same across all year levels (3, 5, 7 and 9).

Key to the validity of this study is whether students were able to demonstrate their writing ability via the computer. 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. 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. Cognitive interviews also revealed that students appreciated the opportunity to edit their work online. Finally, psychometric analyses confirmed that the underlying writing scales performed in a similar manner to their paper-based analogues. Taken together these results indicate that the construct validity of the writing assessment was maintained in the digital environment.

The purpose of the research presented in this paper is to explore the capacity of automated essay scoring systems to replicate human marking of NAPLAN persuasive essays using the persuasive writing rubric. The paper is organised as follows:

- Description and brief research summary of automated scoring of writing
- Method
- Results
- Discussion

[1] For a full summary of NASOP and the findings of all studies released to date, go to the ACARA web site: http://www.acara.edu.au/assessment/research.html

Description and brief research summary of automated scoring of writing[2]

Automated scoring of writing uses computer algorithms designed to emulate human scoring. This is achieved by extracting linguistic features from essays and then using machine learning and modelling to establish a correspondence between these features and essays scores based on a sample of essays that have been scored by human markers. These processes are iterative and require replication to achieve the most optimal solutions for scoring of essays. Consequently, automated scoring solutions require training (primarily through the use of human-marked scripts) before they can be used to mark student essays in a live test administration.

First generation automated scoring solutions developed in the mid to late 20th century used limited language processing methods and algorithms (e.g. word counts, grammar/spell checks) to extract and evaluate lexical and syntactic properties of essays that served as proxies of the underlying writing ability (e.g. Page 1966, 1994; Ajay, Tillett & Page 1973). Regression analyses were then typically applied to generate an essay score. However, over the past two decades, significant developments and advancements in lexical analyses and knowledge representation have extended the coverage of modern automated scoring models to semantic information as well as greatly increased the range of linguistic features that can be extracted from essays.

Latent semantic analyses (LSA), (Landauer & Dumais, 1997; Foltz, Kintsch & Landauer, 1998) and similar semantic content analysis and representation models have been successfully integrated in the automated essays scoring solutions (e.g. Foltz, Laham & Landauer, 1999; Attali & Burstein, 2006). Such integration extended the range of features that can be extracted and used in automated scoring to cover the meaning of words, sentences and paragraphs used in an essay. These models are constructed using a large structured set of texts ("lexical corpora") and thus provide an external frame of reference for extraction and assessment of semantic features in essay autoscoring. LSA uses hundreds of syntactic, lexical and semantic dimensions to represent meaning of essays. Such complex statistical modelling has been enabled by rapid developments in computer

[2] To assist the reader, the following definitions are provided:
Syntax: set of rules, principles, and processes that govern the structure of sentences in a given language
Lexical: of or relating to words or the vocabulary of a language as distinguished from its grammar and construction (i.e. syntax).
Semantics: study of the meanings of words and phrases in language

capacity and power and advances in artificial intelligence theory. Recent developments have also allowed the use of sophisticated, comprehensive methods of machine learning and modelling to establish the connection between lexical and semantic features of essays and their scores.

NAPLAN analytical marking rubrics are explicitly designed to assess differential and combined contributions of lexical and semantic features in writing. The persuasive marking rubric, for example, has criteria that target lexical properties of essays (sentence structure, paragraphing, punctuation and spelling), criteria that target semantic properties of essays (audience and ideas) and criteria that explicitly target successful interplay of lexical and semantic features of writing (test structures and cohesion). Therefore, the current approach to marking of NAPLAN writing is well positioned to utilise the advancements in the field of automated scoring models. These solutions have shown increased sensitivity, reliability, and enable more comprehensive investigation of the validity of automated assessment of writing (see Deane & Quinlan, 2010; Shermis & Bursten, 2013).

A significant growing body of literature confirms that automated essay scoring solutions meet, and in many instances surpass, the quality of human markers (e.g. Page, Poggio & Keith, 1997; Williamson, Bejar & Hone, 1999; Landauer, Laham & Foltz, 2003; Page, 2003; Attali, 2004; Rudner, Garcia & Welch, 2006; Wohlpart, Lindsey & Rademacher, 2008). In 2012, the Hewlett Foundation sponsored a seminal comparative study in the effectiveness of existing commercial automated scoring solutions of essays (Shermis & Hamner, 2012; Shermis & Hamner, 2013). Eight vendors and one university participated in the comparative study. A sample of eight different prompts were used in this study. For each prompt, essays were randomly divided into a training and test sample. Vendors received scores from two human markers only for the training set (some essays also had a third consensus mark). The training set was used to prepare automated essay scoring engines for marking of the student essays. The test set was used as part of a blind test for the score model predictions and to calculate scoring engine performance for a public competition (Shermis & Hamner, 2012).

The Hewlett competition demonstrated that automated engines were able to match or exceed the consistency of human marking. The rate of agreement was higher between any of the automated scoring engines and human markers than that between the two human markers. As with human markers there was variability across vendor performance, a few vendors scored well across all of the prompts while others performed better with certain types of essays. In a second open source competition the three top algorithms

continued to match or exceed human performance with respect to agreement between human and automated scores (Shermis & Hamner, 2013).

NAPLAN persuasive writing tasks require students to draw on personal knowledge and experience to complete their essay. Quality of writing is assessed using an analytic persuasive writing marking guide, consisting of a rubric containing ten marking criteria[3] and annotated sample scripts. The scoring system used in the rubric reflects the weights to be assigned to each component. Consequently, to evaluate the feasibility of automated scoring solutions for NAPLAN online writing tests, it was necessary to obtain information about the performance of automated scoring engines both at the criteria and total rubric score level for each scoring engine.

The following research question was addressed in the study:

> *What is the congruence and reliability of human marking of online NAPLAN persuasive essays at the criterion and total score levels?*

Method

Participants and materials

Using an objective procurement process, automated essay scoring services from four different vendors were obtained for this study, allowing this research to examine the variability of the different automated scoring solutions across vendors.

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. These essays were all typed in by students in the online test delivery system and double marked by two separate groups of markers. Markers reported that online essays matched typical handwritten essays in terms of the length and quality of writing. The descriptive analyses confirmed such observations showing that the average essay length was 118.1, 229.6, 342.4 and 371.1 words for years 3, 5, 7 and 9 respectively and that the median raw score awarded to students was 19.2,

[3] The NAPLAN writing test assesses student performance against ten criteria:

- audience
- text structure
- ideas
- character and setting (narrative); persuasive devices (persuasive)
- vocabulary
- cohesion
- paragraphing
- sentence structure
- punctuation
- spelling

26.0, 30.8 and 33.3 for years 3, 5, 7 and 9 respectively. The scripts were provided in HyperText Markup Language (HTML) format to the contractors. The essays were randomly assigned to one of two groups:

- training and validation essays (N=1014). Contractors were provided with the two sets of human marker scores for each essay. This set was delivered to the contractors with a main set (n=674 essays) and a smaller set that could be used for testing (n=340), again using random assignment of scores.
- test essays (N=339). Contractors were not provided with any marking data for these essays.

Vendors

Four vendors were independently engaged to score the NAPLAN persuasive essays. Each vendor had participated in the Shermis and Hamner (2012) study and is well established in the field. The four engines also represent a good cross-section of approaches and methods for automated assessment of writing.

- Measurement Incorporated provided the Project Essays Grader (PEG) system, which uses lexical and syntactical features in essay marking.
- Pearson provided the Intelligent Essay Assessor (IEA), which uses LSA semantic text analysis. The IEA also takes into account the syntactical features of the writing.
- Pacific Metrics provided Constructed-Response Automated Scoring Engine (CRASE), which uses natural language processing and machine learning models to produce essay scores.
- MetaMetrics provided the Lexile Writing Analyser, which is a scoring engine that does not require any training as it relies on the Lexile scale, which is a proprietary measure of text complexity. This measure of text complexity is developed using syntactic and semantic features of texts.

The four engines also differ in the approaches for extraction of key features required to predict and award scores for each of the criteria.

Procedure

ACARA provided the training and validation set to all of the vendors and then after the vendors had completed system preparation, the test essays were provided. Vendors completed the scoring and provided ACARA with a research report outlining the methods used in their investigation and its key outcomes. Each vendor selected the method for analysing and displaying the results of their analyses (see Results section below). To

foster comparison across models, ACARA provided further analyses of the results submitted by the vendors in order to determine the variability across the four automated scoring solutions. The purpose of this analysis was not to rank models but simply to summarise their performance and enable final interpretation of the study results.

Results

Analyses and results of the evaluation of training and validation stages for each of the automated scoring solutions are detailed in the technical report companion piece to this study. In this section, only the overview of the approaches used to provide criteria scores and key findings will be presented. The comparison across automated scoring solutions performance is also presented.

PEG

To provide marks for NAPLAN persuasive writing PEG constructed a model, which focused directly on the 10 NAPLAN persuasive criteria.

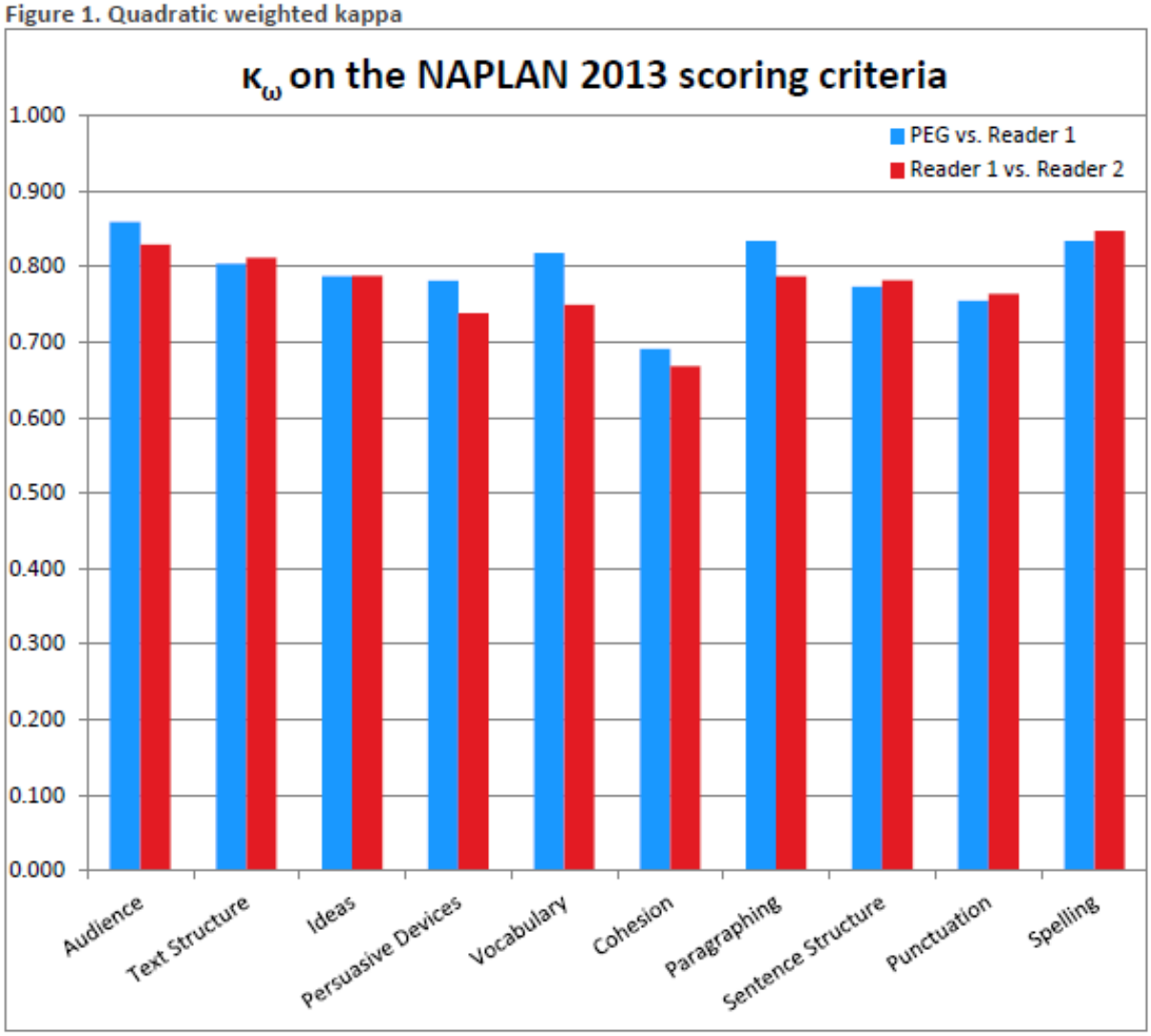
The implemented model identified a set of explicit syntactical and lexical features relevant for each of the criteria. In addition to these explicit features for each of the criteria, PEG identified and extracted implicit patterns and features that are also used to predict the final score for each of the criteria.

Results showed that taking into account a range of measures including quadratic weighted kappa[4], Pearson's r, perfect and adjacent agreement – PEG predicted scores were statistically and substantively equivalent to those provided through human scoring. (See Figure 1 for the summary based on quadratic weighted kappa).

[4] Quadratic Cohen's kappa is a weighted kappa statistic that takes into account the magnitude of the disagreement between pairs of markers. Weighted kappa statistics are appropriate for measuring markers agreement for ordinal data. Quadratic or squared weights use weights that are proportional to the square of the number of categories apart.

Figure 1

Figure 1. Quadratic weighted kappa



IEA

IEA has the capability to score different traits in writing by choosing and weighting different combinations of lexical and semantic features. Some of these existing traits directly match NAPLAN criteria and others were closely related.

The results from the analyses show that IEA was able to closely match, and in some cases, exceed the agreement rates of the human scorers for most of the criteria with slight decrements in performance for the criteria that use structural or paragraphing elements. (See Table 1 for correlations, percentage of exact matches and percentage of adjacent matches between human-to-human, human 1 to IEA and human 2 to IEA).

Table 1. Correlations, percentage of exact matches and percentage of adjacent matches

Trait	H-H Cor	H-H Exact	H-H Adj	IEA-H Cor	IEA-H Exact	IEA-H Adj	h1 to IEA Cor	h1 to IEA Exact	h1 to IEA Adj	h2 to IEA Cor	h2 to IEA Exact	h2 to IEA Adj
1	0.82	61.2	97.6	0.90	61.8	99.7	0.87	62.1	99.1	0.85	60.6	98.8
2	0.81	68.8	99.7	0.86	65.9	99.1	0.83	67.6	98.8	0.81	62.4	99.7
3	0.78	62.4	99.1	0.89	65.9	99.4	0.83	64.1	99.7	0.85	68.2	99.4
4	0.73	55.6	98.5	0.87	64.1	99.4	0.80	61.8	99.4	0.82	63.2	99.7
5	0.76	64.4	97.6	0.86	65.9	99.7	0.80	62.1	98.5	0.81	65.6	98.8
6	0.71	71.8	99.7	0.80	73.5	100	0.74	72.1	100	0.74	70.0	100
7	0.83	69.7	99.4	0.82	57.6	98.5	0.78	55.6	98.5	0.79	60.0	98.5
8	0.78	58.5	95.9	0.85	60.6	97.9	0.81	59.7	98.5	0.80	57.1	96.2
9	0.76	60.6	97.9	0.84	60.0	99.1	0.78	55.0	98.5	0.79	63.2	98.2
10	0.84	66.2	98.8	0.92	67.6	99.4	0.88	65.6	99.1	0.88	67.9	99.4

CRASE

To prepare the system for the marking of criteria scores, CRASE mapped its different feature extraction capabilities against each criterion included in the persuasive rubric. Some of these extraction features are based on the syntactic or lexical properties of written text and others are based on the machine learning models only. The results showed that CRASE was able to exceed the quadratic weighted kappa for all criteria except for paragraphing, as shown in Table 2.

Table 2. Quadratic weighted Kappa agreement rates

	Exact Agreement			Kappa			Quadratic Weighted Kappa		
Marking Criterion	Both Exceed	One Exceed	None Exceed	Both Exceed	One Exceed	None Exceed	Both Exceed	One Exceed	None Exceed
Audience	✓			✓			✓		
Text structure	✓			✓			✓		
Ideas	✓			✓			✓		
Persuasive devices	✓			✓			✓		
Vocabulary	✓			✓			✓		
Cohesion		✓			✓			✓	
Paragraphing			✓			✓			✓
Sentence structure	✓			✓			✓		
Punctuation		✓			✓			✓	
Spelling	✓			✓			✓		

Note. Both Exceed=CRASE-H1 and CRASE-H2 agreement values both exceed H1-H2 agreement value. One Exceed=One of CRASE-H1 or CRASE-H2 agreement values exceeds H1-H2 agreement value. None Exceed= None of CRASE-H1 or CRASE-H2 agreement values exceeds H1-H2

Lexile Writing Analyser

Lexile Writing Analyser implemented a top-down approach in which the total score on an interval Lexile scale was determined for each essay. The machine learning processes were then used to transform such interval measures into an ordinal score for each rubric criterion. The outcome shows that the quadratic weighted kappa reached a satisfactory level for most of the criteria marks (see Table 3).

Table 3. Quadratic weighted Kappa

Machine Predicted Rating (Columns AM through AX)	Human Rating (Columns C through Y)	Quadratic Weighted Kappa Statistic
LWA Rating	Average Total Score	0.8828
Total Rating	Average Total Score	0.9190
Audience Rating	Average Audience	0.8548
Text Structure Rating	Average Text Structure	0.8283
Ideas Rating	Average Ideas	0.8320
Persuasive Devices Rating	Average Persuasive Devices	0.8180
Vocabulary Rating	Average Vocabulary	0.8326
Cohesion Rating	Average Cohesion	0.7411
Paragraphing Rating	Average Paragraphing	0.8094
Sentence Structure Rating	Average Sentence Structure	0.8070
Punctuation Rating	Average Punctuation	0.7146
Spelling Rating	Average Spelling	0.8574

Summary Analyses

As it can be seen in Table 4 below, overall results of the test set of essays confirm that all automated essay scoring solutions achieved the level of agreement with human markers which was observed between the human markers on all criteria. Table 4 also provides an estimate of 95% confidence interval around each of the correlation coefficients. All the observed differences are well within overlapping confidence intervals indicating that these differences are not statistically significant.

Table 4

	Human markers		Automated essay scoring system			
		Marker 2	AES 1	AES 2	AES 3	AES 4
Audience	Marker 1	0.83	0.82	0.85	0.81	0.86
		(0.79-0.87)	(0.78-0.85)	(0.82-0.88)	(0.77-0.84)	(0.84-0.89)
	Marker 2		0.85	0.86	0.83	0.87
			(0.82-0.88)	(0.83-0.89)	(0.79-0.85)	(0.85-0.90)
Text structure	Marker 1	0.81	0.80	0.80	0.78	0.82
		(0.78-0.85)	(0.75-0.83)	(0.76-0.83)	(0.73-0.82)	(0.78-0.86)
	Marker 2		0.83	0.81	0.81	0.84
			(0.80-0.86)	(0.77-0.84)	(0.78-0.84)	(0.81-0.87)
ideas	Marker 1	0.79	0.78	0.78	0.75	0.82
		(0.74-0.83)	(0.73-0.81)	(0.74-0.82)	(0.70-0.79)	(0.77-0.85)
	Marker 2		0.81	0.81	0.77	0.82
			(0.77-0.84)	(0.76-0.85)	(0.72-0.81)	(0.78-0.85)
Persuasive devices	Marker 1	0.74	0.77	0.78	0.75	0.79
		(0.69-0.79)	(0.72-0.81)	(0.74-0.82)	(0.70-0.79)	(0.75-0.83)
	Marker 2		0.81	0.8	0.76	0.8
			(0.77-0.84)	(0.76-0.83)	(0.71-0.79)	(0.76-0.84)
Vocabulary	Marker 1	0.75	0.78	0.81	0.77	0.82
		(0.70-0.80)	(0.74-0.83)	(0.77-0.85)	(0.72-0.81)	(0.77-0.85)
	Marker 2		0.81	0.84	0.78	0.86
			(0.76-0.84)	(0.80-0.87)	(0.74-0.82)	(0.82-0.89)
Cohesion	Marker 1	0.67	0.64	0.67	0.62	0.7
		(0.59-0.73)	(0.57-0.69)	(0.59-0.73)	(0.56-0.68)	(0.63-0.75)
	Marker 2		0.7	0.67	0.63	0.69
			(0.63-0.76)	(0.60-0.74)	(0.56-0.69)	(0.63-0.75)
Paragraphing	Marker 1	0.79	0.8	0.82	0.75	0.82
		(0.74-0.83)	(0.76-0.84)	(0.78-0.85)	(0.71-0.79)	(0.78-0.85)
	Marker 2		0.8	0.8	0.75	0.8
			(0.75-0.84)	(0.76-0.84)	(0.70-0.80)	(0.75-0.84)
Sentence structure	Marker 1	0.78	0.74	0.78	0.72	0.79
		(0.74-0.83)	(0.69-0.79)	(0.73-0.81)	(0.68-0.77)	(0.75-0.83)
	Marker 2		0.77	0.77	0.76	0.81
			(0.71-0.81)	(0.72-0.81)	(0.71-0.80)	(0.77-0.84)
Punctuation	Marker 1	0.77	0.64	0.74	0.73	0.76
		(0.71-0.81)	(0.58-0.70)	(0.69-0.77)	(0.68-0.78)	(0.71-0.80)
	Marker 2		0.68	0.74	0.72	0.77
			(0.63-0.73)	(0.69-0.78)	(0.68-0.77)	(0.72-0.80)
Spelling	Marker 1	0.85	0.82	0.83	0.81	0.87
		(0.80-0.88)	(0.78-0.85)	(0.80-0.87)	(0.77-0.85)	(0.84-0.89)
	Marker 2		0.82	0.82	0.82	0.86
			(0.77-0.85)	(0.77-0.86)	(0.77-0.86)	(0.82-0.89)

The same findings were observed when quadratic weighted kappa statistics were calculated for the total essay scores, as shown in Table 5.

Table 5. Quadratic weighted Kappa

		Human markers	Automated essay scoring system			
		Marker 2	AES 1	AES 2	AES 3	AES 4
Total score	Marker 1	0.79 (0.63-0.87)	0.74 (0.61-0.83	0.76 (0.62-0.84)	0.77 (0.63-0.85	0.73 (0.56-0.83)
	Marker 2		0.72 (0.54-082)	0.77 (0.64-0.85)	0.75 (0.59-0.84)	0.82 (0.67-0.89)

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

Discussion

The purpose of this study was to investigate whether modern automated essay scoring systems prove to be a feasible solution for marking NAPLAN online writing tasks. The investigation showed that at both the rubric criteria and total score levels, the four marking systems provided satisfactory (equivalent or better) results relative to human marking.

These results replicate previous findings (such as the Hewlett competition and others cited in the introduction). Furthermore, these results extend this already considerable body of research with evidence that automated scoring solutions are capable of handling marking rubrics containing 10 different criteria. It is important to note that evidence for successful autoscoring at the criteria level was obtained from all four automated essay scoring systems, which all use different theoretical and methodological approaches to scoring of lexical and semantic features of writing. This finding bodes well for direct application to NAPLAN writing.

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). Even though there is a great deal of similarity between the narrative and persuasive marking rubric, future research will need to include both types of tasks in order to collect comprehensive evidence for the viability of the automated easy scoring solutions for all NAPLAN writing genres.

While this study focused primarily on the reliability of essay marking, NAPLAN online readiness research will also focus on the validity of writing, in particular construct and consequential validity. In particular, ACARA will examine if the introduction of automated scoring has an effect on the substance and quality of student writing ("construct validity") and how writing is taught in the classroom ("consequential" validity). For example, content experts need to analyse the features of discrepant essays (i.e. essays where these was found to be disagreement between human markers and/or computer markers) to examine where the automated scoring markers may be expected to not perform well; if these conditions overlap where human markers fall short; and how the implementation model will address these circumstances. Williamson (2013) and Attali (2013) offer practical advice on how to bring validity into the centre of all planning, implementation and reporting activities in automated scoring of writing.

Transition of NAPLAN to online delivery will provide a better-targeted assessment, more precise measurements and a faster turnaround of results to students and schools. For the writing domain, this provides the opportunity to use the latest developments in artificial intelligence solutions for scoring of writing essays to provide efficient, immediate feedback. The findings of this study suggest that automated essay scoring is effective enough to play a central role in the move of NAPLAN online. Future planned research will determine the extent to which human marking will be needed either to validate or fully supplement this capacity.

REFERENCES

Ajay, H.B., Tillett, P.I., & Page, E.B. (1973). Analysis of essays by computer (AEC-II). *Final report to the National Centre for Educational Research and Development* (Project No. 80101), p. 231.

Attali, Y. (2004). Exploring the feedback and revision features of criterion*. Journal of Second Language Writing, 14*, 191-205.

Attali, Y. (2013). Validity and reliability of 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 R v. 2. *The Journal of Technology, Learning and Assessment, 4*(3).

Deane, P. & Quinlan, T. (2010). What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research*, *2*(2), 151-177.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual Coherence with Latent Semantic Analysis. *Discourse Processes, 25,* 285-307.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. *In proceedings of World Conference on World Media and Technology.*

Landauer, T. K. & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis, theory of acquisition, induction and representation of knowledge. *Psychological Revive, 104* (2), 211-240.

Landauer, T. K., Foltz, P. W. & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259–284.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48*, 238–243.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62*, 127–142.

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Page, E. B., Poggio, J. P., & Keith, T. Z. (1997). Computer analysis of student essays: Finding trait differences in the student profile. *AERA/NCME Symposium on Grading Essays by Computer.*

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric™ essay scoring system. *Journal of Technology, Learning, and Assessment, 4*(4).

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

Shermis, M. D., & Hamner, B. (2012). *Contrasting State-of-the-Art Automated Scoring of Essays:*
Analysis. Paper presented at the annual meeting of the National Council of Measurement in
Education, Vancouver, BC, Canada. Retrieved from: http://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. 313-346). New York, NY: Routledge.

Williamson, D.M. (2013). Probable cause: Developing warrants for automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 153-180). New York, NY: Routledge.

Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). Mental model's comparison of automated and human scoring. *Journal of Educational Measurement, 36*, 158-184.

Wohlpart, A. J., Lindsay, C., & Rademacher, C. (2008). The Reliability of computer software to score essays: Innovations in a humanities course. *Computers and Composition*, *25*, 203-223