Trial of Automated Essay Scoring:
new directions for national assessment in Australia. *

Bronwyn Davies and Tracey Gralton *Benchmarking and Educational Measurement Unit*

BACKGROUND AND RATIONALE

Automated Essay Scoring

Automated Essay Scoring (AES) is the use of a machine to score student essays. Compared to traditional marking, the features of AES include immediate delivery of scores and feedback; customized tasks and scoring systems; consistent and bias-free scoring; and reduction of time and costs of marking by teachers or professionals. Studies testing agreement rates on a "true" score between humans and computers consistently show that AES delivers higher congruence than human scoring. AES requires scripts to be in electronic format, either typed by the student using the software tools of the chosen AES system, or transcribed to typescript from the students' handwriting.

Purpose of the trial

The purposes of the trial were to:

- assess the feasibility and reliability of AES in Australia in a large-scale testing context;
- ascertain attitudes of education stakeholders about AES; and
- ascertain whether human markers are influenced by the mode of script; typed or handwritten.

Research Questions

The research questions were:

- How do human and machine scores compare?
- What degrees of reliability are obtained?
- Is there any discernible difference in the scores of handwritten and typed scripts?
- Is proficiency in writing influenced by the mode of production ie handwritten vs typed?
- What are the perceptions of users and stakeholders?

METHODOLOGY

Participants

2257 students were randomly selected across Years 5, 7 and 9 from a range of schools and settings in Western Australia, South Australia and Queensland. Schools were selected on the basis of their sector and socio-economic status (SES). For practical reasons only metropolitan schools were involved. Year 3 students were excluded from the trial as it was considered that students' computer skills may not be sufficiently developed at this stage for taking the writing task on-line. It is acknowledged that these factors are limitations of the trial, and must be considered in further work.

Sample

Due to the variety of schools involved, the large number of participants in the trial, and the random allocation of students within classes to tasks, it can be assumed that the ability levels of both groups of students were comparable.

Timing

Trialling took place at the schools' discretion between 11th and 25th August 2008.

Materials

The AES system used for this trial was *IntelliMetric*®, the scoring engine operated by Vantage Learning. All participating schools were required to support on-line participation

with a sufficient number of computers and suitable operating platforms and software. The requirements were not in excess of the normal provisions in most schools. Technical support was made available to ensure that systems were operational and that AES software could be accessed. All students were provided with a coloured stimulus text explicating the prompt and the task; a planning sheet and in the case of students taking the pan and paper task, an answer booklet.

Task

Students within each class in the trial were randomly allocated by BEMU either to write their script by hand or to produce it on-line. All students completed the same narrative writing task, with half submitting traditional handwritten papers and half submitting typed electronic scripts. This task was administered under controlled conditions, similar to standards required for state and national tests.

The task required all students to write a narrative based on the topic "*Trapped*". Supervising teachers read a short script outlining some ideas and expectations for completing the task, and instructed students about exam conditions and time allowed. The coloured stimulus text included these instructions together with some images designed to stimulate thinking about the topic.

Data collection

The students' demographic data including SES information; their access to, and experience with computers at home and school; and their level of skills training was collected.

Marking

All scripts were marked against a rubric very similar to that used in NAPLAN 2008 in which application all markers were trained. All markers were highly experienced and representative of the best markers available for jurisdictional or national marking programs. All markers completed 16 control scripts to ensure that they were adhering to the rubric.

Script pool

The sample comprised 2203 scripts written by students in Years 5, 7 and 9. Of those 1053 scripts were typed on-line by the students and 1150 scripts were handwritten. The handwritten scripts were transcribed to typescript for machine marking.

Marking design

The marking process was designed to ensure that there would be sufficient data to answer the research questions. All scripts in their typed version were scored by humans and by *IntelliMetric*®. All pen and paper scripts were scored in the handwritten and transcribed versions by humans. A subset of the transcribed scripts (539) was marked for a second time by humans. In order to measure whether a marker was influenced by the mode of the essay, each marker scored a number of scripts in both the handwritten and typed version. A further subset of 63 scripts was marked by five humans to determine the "true" score.

Attitudinal survey

Following the test, all principals, teachers and students involved in the project were surveyed on their perceptions and attitudes to AES. Focus groups were conducted with students.

ANALYSIS

The analysis consisted of two parts. In the first phase, different methods of marking were investigated using descriptive statistics, agreement rates and Rasch analysis. This forms Section 1 below.

***Descriptive statistics*:** The overall means, standard deviations, standard errors and spread of the human scores and *IntelliMetric*® scores were calculated. Correlations between all combinations of marking type and script mode were also calculated. Matched pairs t-tests were carried out to determine the statistical significance of the various calculations.
***Agreement Rates*:** Rates of agreement between human scores and machine scores with a "true" score for a set of 63 scripts were calculated and compared. Scores were designated as exact if the machine or human score was the same as the "true" score. Scores were designated as discrepant if there was a difference of one or more points.
***Scoring variations*:** Variation among scores allocated to individual scripts by different marking methods was examined.
***Rasch Analysis*:** RUMM 2020 was used to examine the level of fit of the results from both modes of marking to the Rasch model, to determine the validity of the test.

In the second phase of analysis, the different modes of production employed by students in the assessment (pen and paper vs on-line) were investigated. Comparisons between the scores for different modes of production were carried out, and investigations included the analysis of sub groups of students. Qualitative data was collected to investigate the background and perceptions of stakeholders. This forms Section 2 below.

SECTION 1

RESULTS

Descriptive Statistics

The first element of the analysis aimed to answer the question "How do human and machine scores compare?" In order to do this all the handwritten and on-line responses from students were randomly allocated to human markers and scored. The handwritten scripts were then transcribed to typescript to enable machine marking, and together with the on-line scripts were scored by *IntelliMetric*®. A number of other marking combinations were instituted to investigate the functioning of different marking methods in more detail, so that overall results are underpinned by knowledge of how the marking operated for different variables. When the mean scores and standard deviations for all 2203 scripts (on-line and transcribed versions of the handwritten scripts) were compared (Table 1) it was observed that the mean for the machine scores was 1.4 raw score points higher and the standard deviation was 0.6 points lower than the human scores. There was a slightly greater range of scores in human marking compared to the machine.

Script set	Scoring	N	Mean score	Std. Dev	Std. error mean	Min score	Max score
All scripts: TS and OL	Human score	2203	23.9	6.0	.129	7	44
All scripts: TS and OL	Machine score	2203	25.3	5.4	.115	10	41

Table 1: Comparison of human scores and machine scores for the entire script set.

As the study involved two modes of production (handwritten and on-line) and two scoring methods (human and machine), comparisons of various subsets of the student responses have been conducted to investigate the effect of different variables. The following tables (2-8) present the information of these paired sets. Table 2 shows the differences between traditional human scoring of handwritten scripts and the machine score for those same scripts after transcription to typescript. This is an important comparison, as it takes the usual mode and method of marking, and compares it to a mechanised process that assesses an altered artefact (the transcribed script). The mean for the machine-scored transcribed scripts was 1.0 raw score points higher and the standard deviation was 0.7 points lower than the

human scores for the handwritten versions of these scripts.

Script set	Scoring	N	Mean score	Std. Dev	Std. error mean	Min score	Max score
HW	Human score	1150	24.5	5.6	.165	6	43
TS	Machine score	1150	25.5	4.9	.146	10	41

Table 2: Comparison of human-scored handwritten scripts and machine-scored transcribed scripts.

To triangulate and better understand this result, and control for the differences between handwritten and transcribed scripts, students' transcribed writing was scored by randomly allocating the transcribed version of the scripts to human markers. Under this arrangement the comparison of human and machine scoring shown in Table 3 below is based on identical scripts, i.e. the transcribed versions of handwritten scripts. The results show that the mean of the machine scoring was 1.6 raw score points higher and the standard deviation 0.8 lower than the human scoring.

Script set	Scoring	N	Mean score	Std. Dev	Std. error mean	Min score	Max score
TS	Human score	1150	23.9	5.7	.169	7	44
TS	Machine score	1150	25.5	4.9	.146	10	41

Table 3: Comparison of human-scored transcribed scripts and machine-scored transcribed scripts.

In order to investigate whether markers are affected by the mode in which the scripts are presented, humans scored the same script in both its handwritten and typed version. This is an important comparison because the machine cannot mark handwriting, and transcription introduces another variable so this provides baseline information. Table 4 shows that the mean score of the handwritten scripts was 0.6 points higher and the standard deviation 0.1 points lower than the transcribed scripts.

Script set	Scoring	N	Mean score	Std. Dev	Std. error mean	Min score	Max score
HW	Human score	1150	24.5	5.6	.165	6	43
TS	Human score	1150	23.9	5.7	.169	7	44

Table 4: Comparison of human-scored handwritten scripts and human-scored transcribed scripts.

The on-line scripts were also treated as a separate set in order to tease out any differences between humans and machine scoring. Table 5 below shows that the mean human score was 1.1 points lower, and the standard deviation 0.5 points higher than the machine score.

Script set	Scoring	N	Mean score	Std. Dev	Std. error mean	Min score	Max score
OL	Human score	1053	24.0	6.3	.197	7	41
OL	Machine score	1053	25.1	5.8	.181	10	40

Table 5: Comparison of human-scored on-line scripts and machine-scored on-line scripts.

A comparison of two human scores on a subset of transcribed scripts was undertaken to investigate reliability of double human marking. The comparison of human scoring shown in

Table 6 below is based on identical scripts, i.e. the transcribed versions of handwritten scripts and shows that the difference in mean score was 0.6 points and there was no difference in the standard deviation.

Script set	Scoring	N	Mean score	Std. Dev	Std. error mean	Min score	Max score
TSa	Human score 1	539	23.3	5.3	.231	7	40
TSa	Human score 2	539	23.9	5.3	.229	7	38

Table 6: Comparison of two human scores of a subset of transcribed scripts

In order to assess the effect of the script mode on markers each marker scored a set of scripts in both the handwritten and transcribed versions. The matched scripts were distributed to markers in two batches with a gap of several weeks between the first and second sets. The timing was within the total marking period, but the gap was sufficient to ensure the markers had no access to previous scores or scripts, and were unlikely to have detailed recall of the text. Table 7 shows the difference in mean score was 0.1, and the standard deviation differed by 0.2.

Script set	Scoring	N	Mean score	Std. Dev	Std. error mean	Min score	Max score
HWa	Human score	539	23.8	5.5	.24	7	40
TSa	Same human score	539	23.9	5.3	.229	7	38

Table 7: Comparison of scores by the same human on matched handwritten and transcribed scripts

A validation subset of 63 randomly chosen scripts (VS) was marked by five humans to determine the "true" score. Comparisons were made of the human score with the "true" score and the machine score with the "true" score. Table 8 shows that there was no difference between the humans and the "true" score, and a difference of 0.1 between the "true" score and the machine. The standard deviation differed by 1.1 between the "true" score and the human, and by 0.2 between the "true" score and the machine.

Script set	Scoring	N	Mean score	Std. Dev	Std. error mean	Min score	Max score
VS	'True" score	63	24.8	7.7	.977	7	41
VS	Human score	63	24.8	8.9	1.1	7	44
VS	Machine score	63	24.9	7.5	.947	10	41

Table 8: Comparison of human and machine scores with a "true" score for subset of 63 scripts

Matched pairs

The t-test determined that the differences in the mean raw scores amongst the many approaches to marking students' scripts as denoted in the different paired groups in this study are relatively small, and at $p<0.05$ statistically significant in a sample size of $n>2000$.

Correlations

The scores for the machine-marked scripts and those marked by humans show a high correlation of 0.8, suggesting that the basic rank ordering of scripts is preserved. All correlations are statistically significant at the p=0.05 level. It should be noted that the correlation coefficients have not been adjusted for measurement error (attenuation). Had they been adjusted, it is estimated that they would have been larger by an order of approximately 10%. Correlations for the subset pairings ranged between 0.67 and 0.829, except for the "true"

score subset where “true” score and machine or human score were 0.9 for both combinations.

Agreement rates by criteria

The degree of reliability obtained by humans compared with machine scores for the individual marking criteria was investigated. A subset of 63 randomly chosen scripts was marked by five humans to determine the “true” score. Table 9 compares the extent of agreement with a “true” score between humans and the machine. Scores were designated as agreeing exactly or disagreeing. Disagreement comprises differences of one or more score points higher or lower than the true score. In eight out of ten criteria the machine had a higher rate of agreement with the true score than did humans.

CRITERIA	HUMAN SCORE vs TRUE SCORE		MACHINE SCORE vs TRUE SCORE		
	EXACT AGREEMENT	DISAGREEMENT	EXACT AGREEMENT	DISAGREEMENT	difference
AUDIENCE	59%	41%	71%	29%	12%
TEXT STRUCTURE	63%	37%	70%	30%	7%
IDEAS	59%	41%	80%	20%	21%
CHAR/SETTING	52%	48%	73%	27%	21%
VOCABULARY	55%	45%	78%	22%	23%
COHESION	65%	35%	73%	27%	8%
PARAGRAPHING	65%	35%	60%	40%	5%
SENTENCE STRUC	55%	45%	73%	27%	18%
PUNCTUATION	71%	29%	64%	36%	5%
SPELLING	66%	34%	67%	33%	1%

Table 9: Comparison of agreement rates between human and machine with a “true” score for each criterion.

Table 10 compares the agreement rates of the two marking methods with the total score. The first comparison designates agreement to be within two score points of the true score; the second designates agreement to be within five score points. The reason for showing two parameters is to represent both a demanding level of agreement (0-2), and also the acceptable range designated in the NAPLAN marking process (0-5).

TOTAL SCORE	HUMAN SCORE vs TRUE SCORE		MACHINE SCORE vs TRUE SCORE		
	AGREEMENT	DISAGREEMENT	AGREEMENT	DISAGREEMENT	difference
0-2 score points	58%	42%	71%	29%	13%
0-5 score points	90%	10%	97%	3%	7%

Table 10: Comparison of agreement rates between human and machine scores and a “true” score for the total.

Score variations

In addition to the overall picture of the functioning of different marking modes, it is necessary to investigate the extent of variation of scores for individual scripts, because this is where inaccuracies could result in students receiving scores that do not reflect their performance. The first three graphs are derived from a subset of 539 scripts that were double marked by humans. Figure 1 shows the differences of total scores for individual scripts between two humans. Figure 2 shows the differences of total scores between a human and the machine. Figure 3 shows the variation of scores allocated to the same script in its handwritten and typed version by the same marker after a three week interval. The same process was carried out for the set of 63 scripts that are considered to have a “true” score in Figures 4 and 5. Differences greater than three points occurred 28% of the time for human markers, compared with 12% for the machine. Some scores allocated to a script differed by as much as 17 points between the human and machine scores for a

transcribed script, and 16 points between two human scores on the handwritten and transcribed versions. The same human re-marking a paper varied by as much as 11 points.

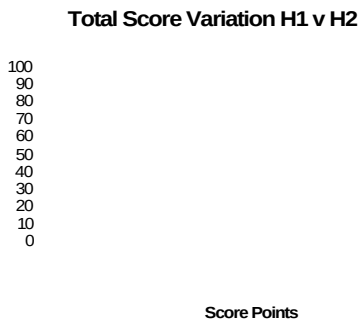

Figure 1

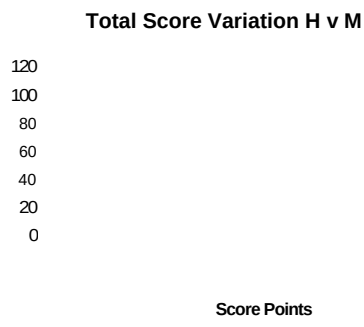

Figure 2

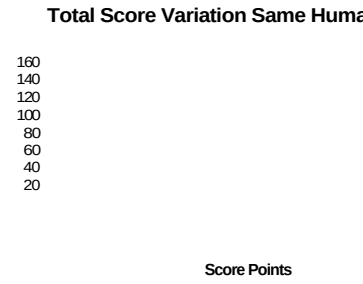

Figure 3

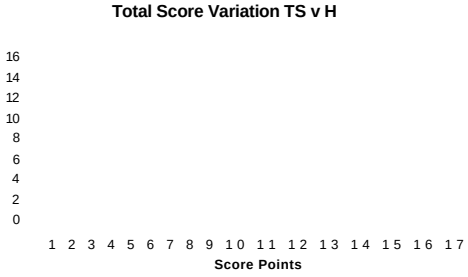

Figure 4: Subset of 63 “true” score scripts

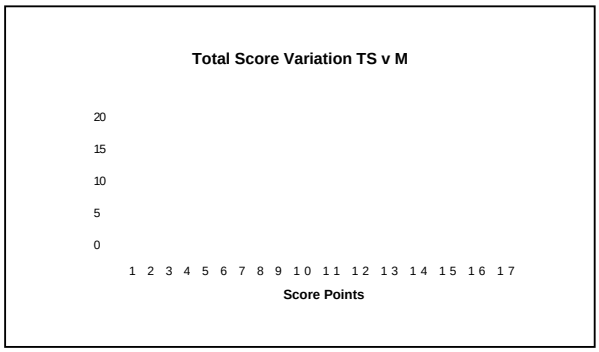

Figure 5: Subset of 63 “true” score scripts

Rasch analysis

The complete set of machine scores and the complete set of human scores were analysed in terms of the Rasch model, using RUMM2020. The reliability for machine scores was 0.96 (Person Separation Index and Cronbach’s Alpha). In terms of targeting the score points were spread over the entire range of abilities. However the range of abilities as reported by the analysis is extremely large (over 45 logits). Although there were no reverse thresholds, paragraphing appears to have some anomalous fit properties, especially in the range 0 – 5 logits, where the observed values appear to be ‘flat’, suggesting under-discrimination. The reliability for the human scored scripts was 0.93, and the range of abilities was spread over 30 logits. The ICC curves show that paragraphing is operating more appropriately in this analysis. A plot of human-scored item locations with the machine-scored locations was not linear, and further investigation of the significance of this phenomenon is needed to discern whether it has any bearing on the validity and reliability of machine scoring.

DISCUSSION

The summary statistics for the whole set of scripts and for various subsets revealed that the mean human score is always slightly lower than the mean machine score and the standard deviation of human scoring is always slightly larger than that of the machine. Human and machine scores therefore were not precisely commensurate although the mean difference of 1.4 raw score points is minimal. The small but statistically significant differences between human and machine scoring may be due to a several factors, such as variation in human marking. The scoring engine was calibrated to a high level of accuracy in an initial study conducted in 2007. It is likely that the machine operates consistently, however human marking may differ from one group or time to another. Human inconsistency has not been explored in depth through such methods as comparison of mean scores of marking centres in different locations and times. However the subset of "true" score scripts showed far smaller differences in mean scores suggesting that the more accurate the human marking the smaller the differences from the machine score. The correlation of 0.8 for human and machine scoring over all scripts is acceptable given that the human marking cannot be considered to be absolutely accurate. By comparison, the correlation of 0.9 is excellent for the "true" score subset where the rank ordering of the scripts is highly accurate

Based on the summary statistics it seems likely that human markers are positively influenced by handwriting, as handwritten versions of scripts received higher mean scores. This may be due to the preconceptions readers have in relation to printed texts which are expected to be polished and error-free; attributes that are not assumed for a handwritten draft.

The slightly compressed results for machine scoring is commonly a feature of comparisons between human and machine marking and this is generally considered to be related to the set of training scripts used to calibrate the essay-scoring machine. The maximum and minimum scores need to be well exemplified to ensure scores are spread properly and these scripts are often difficult to source. In the training set provided to *Vantage Learning* by BEMU there were fewer scripts at the highest score points than is desirable.

The extent of agreement with a true score is better overall for the machine than for humans, whether the tolerance margin was two points or five points. The extent of agreement with the individual criteria scores was also greater for the machine in eight out of the ten criteria, with an average margin of around 10%. This is an important measure as writing assessments are reported at the individual level and while mean scores and correlations are useful population information, the criteria scores will be treated as diagnostic information by teachers.

Although mean scores of all marking and script types are comparable, and correlations are satisfactory, this does not take variation among markers into account. The variation in scores for individual scripts shows that humans differ from each other and the machine; and when compared to a true score, the degree of difference of more than three score points is twice as large for humans than the machine.

The Rasch analyses revealed that the humans and machine appear to be marking slightly different traces within the construct. The results for each type of marking showed good reliability, and a high degree of discrimination. The extremely large range for the machine scores seems to be particularly pronounced at the lower ability range.

SECTION 2

RESULTS

Mode of writing: a comparison of on-line scores and handwritten scores

To determine if mode of writing i.e. on-line or pen and paper, had differentiated effects on student performance, a comparison has been made between those students who completed their essay on-line and those who wrote their essay by hand, by calculating the spread of scores, standard deviations and mean scores.

The data below are based on the machine scores for both sets of scripts. To ensure that any conclusions drawn can be attributed to mode of writing, and not marker variability, the **machine scores** have been used for this analysis.

Table 11 and Figure 6 (below) show the distribution of students' score for students who completed their writing on-line and those who used pen and paper.

	NUMBER OF STUDENTS	% OF STUDENTS	MEAN SCORE	STANDARD DEVIATION	MIN –MAX SCORE
On-line	1053	48%	25.1	5.9	10 - 40
Handwritten	1150	52%	25.2	4.9	10 -41

Table 11: On-line scores v Handwritten Scores: all students

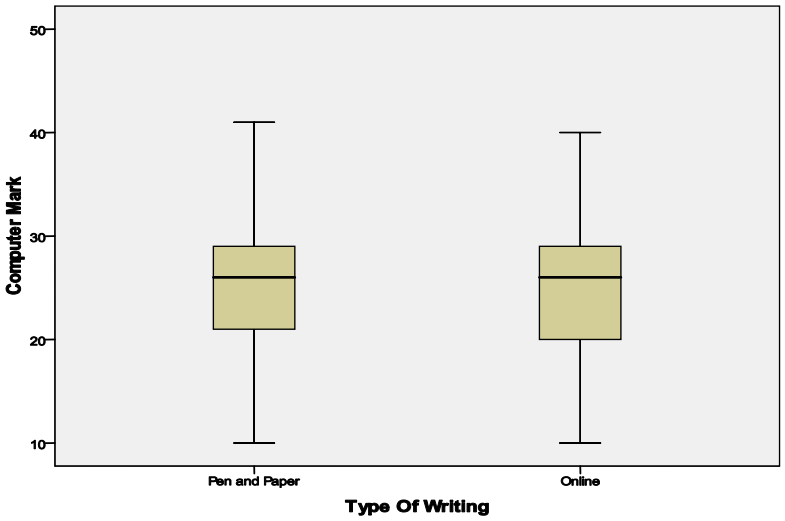

Figure 6: On-line scores v Handwritten Scores: all students

Sub-group analysis

To determine the extent to which the mode of writing exerts an effect on different sub-groups (year level, gender, Aboriginal Torres Strait Islander, Language Background Other Than English), a comprehensive analysis was carried out. The results are shown in Table 12. Although the percentage of students in each sub-group varies, the percentage of students **within** each sub-group who completed their writing by each mode is similar. The number of ATSI students and LBOTE students who participated in the trial was very low and therefore no valid comparisons can be made in regards to their performance by mode.

Table 12 is an overview of all sub-groups. Separate sub-group analyses have been extracted

from this table.

	NO. OF STUDENTS		MEAN SCORE		STANDARD DEVIATION		MIN – MAX SCORE	
	on-line	handwritten	on-line	handwritten	on-line	handwritten	on-line	handwritten
YEAR								
5	323	345	21.6	23.2	4.8	4.5	10 - 35	10 - 35
7	444	496	25.7	25.5	4.5	4.7	11 - 40	12 - 41
9	268	308	28.7	28.1	5.0	4.4	12 - 40	14 - 41
GENDER								
Girl	516	578	26.6	26.2	5.8	5.2	10 - 40	10 - 41
Boy	519	571	23.8	24.7	5.6	4.7	10 - 40	11 - 38
GENDER BY YEAR LEVEL								
YR5 Girls	158	168	22.5	23.7	4.9	4.5	10 - 35	10 - 35
YR5 Boys	165	177	20.8	22.7	4.5	4.5	10 - 34	11 - 34
YR7 Girls	210	243	27.1	25.8	4.5	4.5	11 - 40	13 - 41
YR7 Boys	234	253	25.6	25.2	5.5	4.5	12 - 40	12 - 38
YR9 Girls	148	167	30.2	29.5	4.7	4.3	13 - 40	18 - 41
YR9 Boys	120	141	26.8	26.6	4.8	4.1	12 - 37	14 - 36

Table 12: On-line scores v Handwritten scores: by Sub-Group

Year Level

Table 13 and Figure 7 show the comparison of performance by writing mode for Year Level only.

	NO. OF STUDENTS		MEAN SCORE		STANDARD DEVIATION		MIN – MAX SCORE	
	on-line	handwritten	on-line	handwritten	on-line	handwritten	on-line	handwritten
5	323	345	21.6	23.2	4.8	4.5	10 - 35	10 - 35
7	444	496	25.7	25.5	4.5	4.7	11 - 40	12 - 41
9	268	308	28.7	28.1	5.0	4.4	12 - 40	14 - 41

Table 13: On-Line Scores v Handwritten Scores: By Year Level

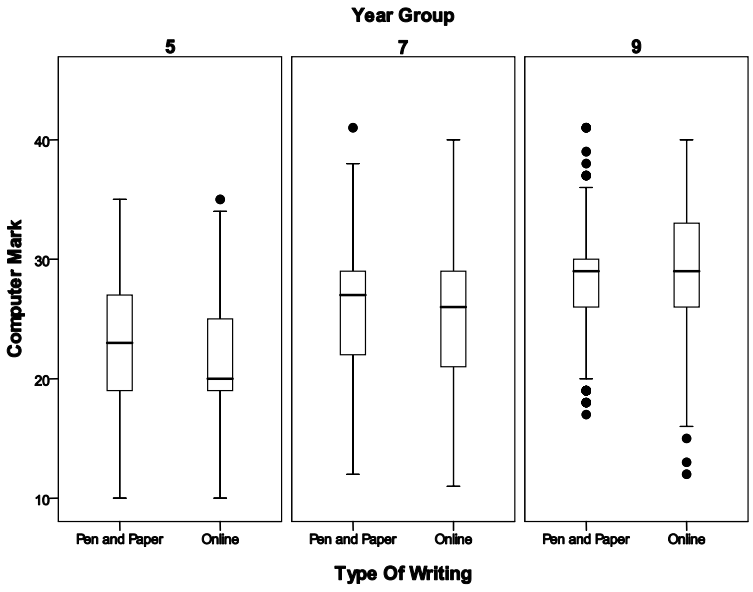

Figure 7: Box and Whisker Plot

Gender

Table 14 and Figure 8 show the comparison of performance by mode for Gender only.

	NO. OF STUDENTS		MEAN SCORE		STANDARD DEVIATION		MIN – MAX SCORE	
	on-line	handwritten	on-line	handwritten	on-line	handwritten	on-line	handwritten
GIRL	516	578	26.6	26.2	5.8	5.2	10 - 40	10 - 41
BOY	519	571	23.8	24.7	5.6	4.7	10 - 40	11 - 38

Table 14: On-Line Scores v Handwritten Scores: By Gender

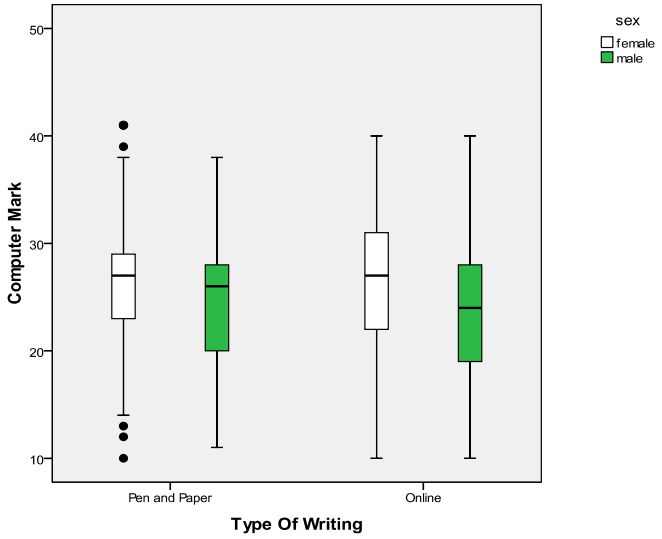

Figure 8: On-line scores v Handwritten Scores: by Gender

DISCUSSION

Group

At the group level the mean score as well as minimum and maximum score for the two modes of writing is almost identical. The standard deviation for on-line scripts is one score point more than for scripts completed by pen and paper.

Year Level

Year 5 – The mean score for scripts completed using pen and paper is higher than for those completed on-line. The distribution for handwritten scripts is more than for on-line scripts.
Year 7 and 9 – The mean score for scripts competed on-line is very similar to those competed using pen and paper. The distribution for on-line scripts is a lot more even at Years 7 and 9 than for those completed by hand.

Gender

Boys – The mean score for boys who used pen and paper is almost one score point higher than for those who wrote on-line. However the highest on-line score is 2 score points higher than for those competed using pen and paper.

Girls – The minimum and maximum score, mean score and standard deviation for students who hand wrote their essays, closely reflected those for scripts completed on-line.

Both – The spread of scores for both boys and girls who completed their essays on-line is more even than the scores for students who used pen and paper for completion of essay.

RESULTS

Word counts: a comparison of on-line essays and handwritten essays

Discussions with stake-holders pre and post testing revealed a common perception that some students are disadvantaged by having to write their essays on-line. It was thought that students who have limited keyboard skills would not be able to write as much as they could if they were using pen and paper, and therefore not perform as highly. It could be argued however that students may be able to write more on-line given that students are becoming more immersed in technology through the internet and web-based education programs and more increasingly write assignments using word processing. Therefore, students adept at using a computer may actually be advantaged by being able to write directly on-line.

A word count of all scripts, both hand-written and on-line was generated and analysed to determine any effect mode has on length of text.

The results are shown below in Table 15.

	NO. OF STUDENTS		MEAN WORD COUNT		MIN - MAX WORD COUNTS	
	on-line	**handwritten**	**on-line**	**handwritten**	**on-line**	**handwritten**
GROUP	1035	1149	351	304	14 - 1410	18 - 817
YEAR						
5	323	345	239	251	31 - 701	18 - 590
7	444	496	373	309	14 - 1410	26 - 817
9	268	308	455	352	39 - 1212	42 - 759
GENDER						
Girls	516	578	394	335	31 - 1212	18 - 817
Boys	519	571	311	272	14 - 1410	26 - 759
GENDER BY YEAR LEVEL						
YR5 Girls	158	168	258	281	41 - 701	18 - 590
YR5 Boys	165	177	222	221	31 - 675	57 - 491
YR7 Girls	210	243	416	331	31 - 935	68 - 817
YR7 Boys	234	253	373	288	14 - 1410	26 - 626
YR9 Girls	148	167	509	392	39 - 1212	80 - 714
YR9 Boys	120	141	390	305	62 - 824	42 - 759

Table 15: Comparison of Word Counts by Sub-Group

DISCUSSION

Group

As a group the mean word count for students who completed their writing on-line was higher by 46 words (15%) than those who completed their essay using pen and paper.

Year Level

Years 7 and 9 students who wrote directly on-line wrote considerably more than their counterparts who used pen and paper, i.e. 21% more at Year 7 and 29% more at Year 9. Although the mean word count for Year 5 on-line scripts was lower than for handwritten scripts, the minimum and maximum marks were significantly larger, indicating that even the slowest typist wrote nearly twice as many words as the slowest writer in the pen and paper test.

Gender

When analysing the word count by gender, both boys and girls who completed their essay on-line wrote more than those using pen and paper, i.e. 18% more for girls and 14% more for boys.

Correlations: experience with computers and on-line word count

In order to determine whether students' experience with computers contributes to the amount they were able to write on-line, correlations between the students' responses provided via an on-line survey and the word count of their essay were calculated.
The correlation between computer use, years of use, frequency, nature of use and the number of words a student wrote on-line was negligible. This indicates that students are neither advantaged nor disadvantaged by having access to computers.

Computer Use	Spearman Correlation
Number of Years Using a Computer	-0.03
Frequency of Computer Use at Home	0.04
Frequency of Computer Use at School	0.17
Frequency of typing assignments/essays	0.02

Table 16: Spearman Correlations between Word Count and Computer Use

Stakeholder perceptions

During the trial of Automated Essay Scoring, both teachers and students were surveyed in a variety of ways, including written surveys, on-line surveys and focus group discussions, in regards to their perceptions of the program. Based on information collected it was shown that:

- the majority of students from all year levels:
  – prefer to type than use pen and paper
  – believe they can write more using word processing than using pen and paper
  – trust that a machine can reliably mark their writing
  – were enthusiastic about writing on-line and receiving instant feedback
- the majority of teachers felt that:
  – their students preferred writing their essay on-line
  – their students would write less on-line than if using pen and paper
  – their students liked receiving instant feedback
  – they were unsure if the machine could reliably mark their students' writing – were unsure if machine scoring was fairer than human marking

KEY FINDINGS

1. Results demonstrate that Automated Essay Scoring produces similar results to scoring by human markers.
2. Results demonstrate that agreement rates with a 'true' score are greater for the machine scores than the human scores.
3. Results indicate that the mode of writing i.e. on-line or pen and paper, had no discernible effect on performance at either the group or sub-group level.
4. Results show that the majority of students are able to write more on-line than using pen and paper.
5. Results indicate that neither students' experience with computers, nor frequency of computer use had any effect on how much they were able to write on-line.
6. Survey results suggest that the majority of students prefer to type their work rather than handwrite. Students believe that they are able to write more on a computer, and that computers can reliably mark their writing.

This research was undertaken for the Performance Measurement and Reporting Taskforce on behalf of all states and territories.