Computer scoring and quality of thought in assessing writing

This paper describes the use of computers to score writing tests, and canvasses the strengths and weaknesses of computer scoring. The paper then reviews, from a distance, some aspects of the development since 1998 of the *e-rater* computer scoring software of the Educational Testing Service. Some of the published attempts to explore the validity of *e-rater* scoring are reviewed, with particular emphasis on the attempt to ground the use of computer scoring on a clear view of writing ability that could give a substantive rationale for the use of such software, and the interpretation of the scores it produces. The different ways in which writing ability can be understood are examined, with particular reference to the content/thought versus form/language dualism. The attempts to have computer software deal with the crucial issue of the quality of thought in writing is discussed. The changes from the first to the second version of *e-rater* are reviewed. The development of a nuanced and sophisticated view of writing ability, and the possible use of such a view as the basis for a specific and circumscribed use of computer scoring in the future is considered.

Computer scoring of writing

Computer scoring of writing (CSW), as it will be called here[1], was first considered in the 1960s, but it was not until the micro computer revolution of the 1980s that such scoring became a realistic proposition. A number of groups and corporations in the United States developed software for scoring essays in the 1990s. In recent years CSW software has contributed to results in such significant, high stakes tests as the Graduate Record Examination (GRE), the Medical College Admissions Test (MCAT) and the Graduate Management Admissions Test (GMAT) which are used for entry to graduate programs in the United States. Computer scoring has been actively considered for use in The National Assessment of Educational Progress (NAEP) in the U.S., and other national and international studies (Horkay, 2005). In high stakes programs like the GRE, MCAT and the GMAT candidates write on a computer, and their work is graded by a human reader and the computer software. Where there is a discrepancy between the human and the computer scores in these testing programs (usually a difference of more than one score point on a six point scale) another human score is used to resolve the difference.

The four major commercially available ‘automated essay scoring systems’ are Project Essay Grade (PEG) of Measurement Inc., Intelligent Essay Assessor (IEA) of Pearson Knowledge Technologies, IntelliMetric of Vantage Learning, and *e-rater* of Educational Testing Service (Keith, 2003).

The proponents of CSW claim that the use of grading software is not only comparatively inexpensive, it is highly reliable in that the machines agree with the grades of humans as much as humans agree with other humans. Any tests where candidates produce electronic ‘scripts’ are of interest to testing companies and educational administrators as candidates for CSW, and as more tests are undertaken using word processors, and on-line, the use of CSW is likely to expand significantly in the near future. Any testing program where candidates produce responses on a computer is a likely candidate for CSW.

As well as the cost advantages, some proponents make significant claims for the positive educational potential of CSW (Y. Attali, & Burstein, J., 2006; Chung, 2003) Some advocates argue that CSW can do more than merely score writing as well as human readers, it is claimed that the software can also offer substantive feedback automatically to students and relieve teachers of the task of responding to all the writing a student produces (Shemis & Burstein, 2003).

The basis of computer scoring

CSW can be undertaken by using a wide variety of the computable characteristics of written language. Some CSW is quite simple, involving not much more than the kind of counts that are used in basic text readability formulas. A basic readability formula, the Flesch, is described and can be used through Microsoft Word. The Flesch formula is a very rudimentary computation of the complexity of written text on the basis of average words per sentence and syllables per word. A more complicated computation is undertaken by the AutoSummary feature in Microsoft Word. Autosummary is not unlike the vocabulary analysis of the content of text undertaken by some CSW. The calculations of the Flesch formula and Autosummary are, of course, completely blind to the meaning of a text, but the proponents of the more complicated CS algorithms claim to compute powerful ‘proxies’ for the semantic content of written language.

Two prominent and sophisticated commercial CSW will be outlined below: IntelliMetric developed and owned by Vantage Learning (Rudner, 2005) and *e-rater* developed and owned by the Educational Test Service (Burstein, 1998; Shemis & Burstein, 2003). The early versions of CS software were 'black boxes' in that the variables assessed were not specified, and the weightings given to different variables were not disclosed. These early versions of CSW were primarily intended for application in testing situations, and they were promoted solely on the basis of their ability to mimic the scores of human readers.

The validity of such 'black box assessments' has been challenged because the meaning and stability of the scores produced was uncertain (Ericsson, 2006). The black box versions of CSW were also criticised because they only produced overall grades. More recent versions of these pieces of software have been re-designed to produce substantive sub-scores that aim to give a kind of formative feedback to students.

The IntelliMetric software of Vantage Learning

The IntelliMetric software (and the first version of *e-rater*) develops topic specific scoring models for particular tasks by analysing the characteristics of scripts that have been marked by a number of humans. A broad array of characteristics is analysed in the 'training scripts' that have been marked by humans so as to construct the typical characteristic for the different grade levels. The values computed for different levels of the training scripts are compared with those of other scripts in an actual scoring. In essence, both IntelliMetric and the first version of *e-rater* compare individual scripts with parameters derived from a group of norming scripts graded by humans.

The IntelliMetric software (Edelblut, 2005) claims to analyse some 400 semantic, syntactic and discourse level features to analyse content and structure. Content includes the breadth of content, and the support for concepts advanced (e.g., vocabulary, concepts, support, elaboration, word choice). Cohesiveness and consistency includes purpose and main idea. The logic of the discourse includes transitional fluidity and relationships among parts of the response (e.g., introduction and conclusion, coordination and subordination, logical structure, logical transitions, sequence of ideas). Structure includes conformity to the conventions of edited American English (e.g., grammar, spelling, capitalisation, sentence completeness, punctuation). Sentence complexity and variety includes usage, readability and subject-verb agreement.

IntelliMetric has no pre-defined rules that are used to score writing samples: the model for scoring emerges from the analysis of the training scripts. An IntelliMetric scoring model is topic specific, and it is shaped by the criteria used for the assessment and the culture of the marking team. If a group of scripts is scored different criteria or by teams working on different assumptions, IntelliMetric might use different parameters to predict the different sets of scores produced.

IntelliMetric uses natural language processing software to parse the training scripts to 'understand' the syntactic and grammatical structure of the language in which they are written. Each sentence is analysed in terms of parts of speech, vocabulary, sentence structure, and concept expression. Several techniques are used to 'make sense of the text' including morphological analysis, spelling recognition, collocation grammar and word boundary detection. A word vocabulary and word concept net are also used to 'form an understanding of the text'.

The analysis of syntactic and discourse characteristics produces an overall score as well as scores for five major domains:

- focus and meaning - cohesiveness and consistency in perspective and main idea;
- organization - logical sequence of ideas and discourse;
- content and development - content breadth, support of theme, elaboration;

- language use and style - word/sentence complexity and variety, tone/voice; and
- mechanics and conventions - adherence to rules of edited American English.

The IntelliMetric software is the basis of an online writing program that assesses students' writing ability and provides suggestions for improvement. It is claimed that this program helps students to:

- produce more writing;
- develop a repertoire of prewriting, drafting, organizing and revising strategies;
- learn to use rubrics for self-guided instruction;
- understand feedback on scoring for 'the five traits of effective writing';
- revise their essays and immediately receive new scores for their revisions; and
- use a Writer's Guide containing activities aimed at improving students' writing.

Vantage Learning also claims that their research has shown that the IntelliMetric software (Vantage, 2004):

- agrees consistently with expert human scoring;
- accurately scores responses across grade levels and different subject areas; and
- shows a strong relationship to other assessments of writing ability.

The strengths of and the opportunities offered by CSW

It has been argued by proponents that CSW:

- is objective and reliable, and produces consistent scores over time;
- is cost-effective quality control in testing programs, especially programs which have single rather than double marking;
- would cut costs in testing programs and encourage more direct testing of writing;
- can offer increased writing practice and substantive feedback to students;
- can motivate students by giving instant feedback to work produced;
- facilitates reworking and reassessment of pieces of writing;
- encourages the testing of writing using computers; and
- would relieve teachers of some of the burden of responding to student writing.

The weaknesses of and the threats posed by CSW

The typical claims that CSW agrees with human scores as much as (or more) than different human scores agree with each other has generally passed without much challenge. The reliability of CSW scores should be subject to more questioning than it has been to date. Most of the claims about the reliability of CSW have been made about tightly constrained, high stakes tests such as GRE, MCAT and GMAT. The writing topics in these tests are specific and circumscribed, and there is reason for thinking that it is these tight topics that enable CSW to mimic human readers. The success of CSW in predicting human scores depends on the extent to which the writing produced by candidates is constrained.

The claims about agreement between CS and human scores make no differentiation between different kinds of human score. Leaving aside differences between individual markers (which are substantial, of course) different testing programs can set out to assess in different ways and on different assumptions. A test of language proficiency like the Test of English as a Foreign Language (TOEFL) or the International English Language Testing System (IELTS) might aim to test something quite different from tests for native speakers like GRE, MCAT or GMAT. The relationship between CS and human scores might be quite different for different testing programs and testing agencies. The human scores produced by different testing programs can differ significantly.

A significant piece of evidence that challenges the claims about the agreement of CSW and human readers has arisen from the NAEP report on *Online Assessment in Mathematics and Writing: Reports from the NAEP Technology-Based Assessment Project* (Sandene, 2005). The independent and rigorous NAEP study offered the following conclusion about the agreement between CSW and human readers:

> *Results showed that the automated scoring of essay responses did not agree with the scores awarded by human readers. The automated scoring produced mean scores that were significantly higher than the mean scores awarded by human readers. Second, the automated scores agreed less frequently with the readers in level than the readers agreed with each other. Finally, the automated scores agreed less with the readers in rank order than the readers agreed with one another.*

In a trial by the author using scripts from a test taken by tertiary aspirants in Canberra Australia, (the ACT Scaling Test) two different pieces of CSW software were unable to match the discrimination between the candidates and the degree of agreement achieved in doing this by human markers (McCurry, 2010). The writing component of this test is a very broad and open prompt that allows candidates to develop a response in different ways. It is arguable that CSW software cannot deal with broad and open writing prompts as well as humans can, and that CSW requires specific and narrow writing tasks (such as those used in the high stakes tests referred to above) to mimic the scores of human readers. The marking of the ACT Scaling Test gives clear priority to assessing quality of thought over language control issues. Markers are encouraged in that test to focus on what candidates are saying rather than how they are using language.

As mentioned above, the validity and the meaning of the scores produced by CSW has been questioned. According to this argument, it is not enough for CSW to mimic human scores, the CSW scores must have substantive meaning. It is claimed that CSW scores have substantive meaning if they are based on a declared model of quality in writing, and the model offers a basis for explicit, substantive feedback. Such criticism seemed to prompt the development of second version of *e-rater* that can be used to offer feedback to students. IntelliMetric continues to base its grading on topic specific models (and continues to promote its product on the basis of agreement with human readers), but it has also developed feedback for students from their CSW. The topic specific scoring of IntelliMetric (and the first version of *e-rater*) has been criticised for producing scores for particular prompts rather than scores for writing ability in general. The second version of *e-rater* discussed below tries to overcome this criticism by using a more or less set scoring model, and by testing the consistency of the scoring across different tasks.

CSW has been criticised because it can be tricked by simply repeating slabs of text or by producing polysyllabic and grammatically correct nonsense (Powers, 2001), but this is not a major problem when both human readers and CSW are used (Monaghan, 2005). A more substantial problem is that CSW will consistently under-rate clear and simple writing that presents sophisticated ideas, and will instead reward syntactically complex and polysyllabic text where the ideas are superficial or even garbled. There is reason to be concerned about the way the CSW software automatically takes complexity of language to be the same as, or a ‘proxy’ for, complexity of thought.

There is a good deal more to be learned about the nature of CSW. There has been no reported study to date of the scripts where humans and machines disagree, and the reasons why they disagree. When human readers and machines substantially disagree about a script, how often does a second reader concur with the first reader or the machine? To what extent does the machine pick up slips made by individual human readers? What do human readers and machines consistently differ about, and why? There may well be a significant pattern in these disagreements. No analysis of the actual relationship of human and machine scores has yet been reported from a high stakes testing program.

It can be argued that CSW:

- offers a distorted view of writing and the writing process;
- reduces quality in writing to the computable characteristics of language;
- assesses linguistic complexity rather than quality of thought or development of ideas;
- turns writing into a mechanical exercise rather than a human communication;
- encourages the production of writing to trick or play to the machine rather than authentic thinking and writing;
- may encourage the spread of testing; and
- may eventually lead to the elimination of the human markers from writing assessment programs.

The development of *e-rater* by the Education Testing Service

CSW has generated a good deal of heat and some light over the last 15 years (Ericsson, 2006). To their credit, the most light has been generated by various researchers at Educational Testing Service (ETS) about their *e-rater* software. The following discussion is based on some of the many published reports about *e-rater* and writing assessment produced over the past 15 years that are readily available of the ETS website.

There would seem to be an interesting history to be written about the seemingly dialectical development in the discussion of CSW in ETS reports. In 1998 Burstein et al. announced the development of *e-rater* with the claim that it showed ‘between 87% and 94% agreement with expert readers' scores, and accuracy comparable to that between two expert readers’ (Burstein, 1998). It was concluded that these results indicated that *e-rater* might be ‘useful as a second reader for high stakes assessments, thus leading to a considerable reduction in essay scoring costs’. *E-rater* was soon being used to replace one of two human readers in such significant, high stakes tests as the GRE, MCAT and the GMAT.

In the same year as Burstein announced the successful development of *e-rater*, Bennett and Bejar of ETS in an article entitled *Validity and Automated Scoring: It’s Not Only the Scoring* (R. Bennett, & Bejar, I., 1998) argued that ‘automated scoring needs to be designed as part of a construct-driven, integrated system’ so that computer testing could be ‘a system consisting of a construct definition, test design, and task design; examinee interface; tutorial; test development tools; automated scoring; and reporting’.

The initial version of *e-rater* was based on what Bennett and Ben-Simon later described as ‘brute empiricism’ (R. E. Bennett, & Ben-Simon, A.. 2005). The first version of *e-rater* was empirical in that it took a broad range of computable features of text and used an optimal combination of them (whatever it was) to give the best possible prediction of the scores given by a particular group of humans for a particular test prompt. Bennett and Ben-Simon, on the other hand, were suggesting that CSW should be ‘construct driven’ rather than pragmatically empirical.

The validity of *e-rater* scoring was addressed by a number of ETS researchers in the years that followed. In 2004 Bennet wrote about 'moving the field forward' by improving the quality of CSW 'through finer control of underlying constructs' (R. Bennett, 2004). He argued that CS needed to be 'grounded in a credible theory of domain proficiency', which was not the case with the first version of *e-rater*. Bennett suggested that the absence of such grounding was an on-going challenge to the credibility and quality of CS research and development. He called for a rigorous and responsible approach to the development of CSW:

> *We can also move the field forward by conducting rigorous scientific research that helps build a strong validity argument; publishing research results in peer-reviewed measurement journals where the quality can be critiqued from a technical perspective; attempting to use competing scoring approaches in combination; employing multiple pieces of evidence to bolster the meaning of automated scores; and, finally, using automated scoring to fundamentally alter the character of large-scale testing in ways that bring lasting educational impact.*

The study by Bennett and Ben-Simon in 2005 about the need for a 'theoretically meaningful' computer scoring system is a landmark in the development of CSW (Ben-Simon, 2007; R. E. Bennett, & Ben-Simon, A.. 2005). In this study Bennett and Ben-Simon developed a 'theoretically driven' method for automatically scoring the writing assessments of the significant and prestigious NAEP. The object of the study was to explore the way writing experts believed different aspects of writing should be weighted to construct an overall score in comparison with the 'purely statistical method' used by *e-rater* and most other CSW software.

Bennett and Ben-Simon used various configurations of *e-rater* to compare the performance of three approaches to CSW. There was the 'brute-empirical approach' in which characteristics of writing were selected and weighted to construct a score solely on the basis of their correlation with human scores. They also developed a 'hybrid approach' in which a fixed set of dimensions which were more closely tied to the characteristics of good writing were used, but the weights for the variables were still largely statistically determined. The third alternative was a 'theoretically driven approach' in which a fixed set of dimensions was weighted according to the judgments of writing experts about the most appropriate contribution of each to an overall score.

The study was building on the development of *e-rater* v2 which differed from the first version of the software in that it could be configured to construct a score on the definite and declared set of dimensions presented in Table 1 (Y. Attali, & Burstein, J., 2006). This version of *e-rater* was a movement away from 'brute empiricism' towards a score that could be related to generally recognised characteristics of good writing.

The research of Bennett and Ben-Simon aimed to explore whether writing experts would agree about the most appropriate weighting for *e-rater* v2 dimensions; the similarity and difference between the scores produced by the three approaches to constructing an overall score from *e-rater*, and whether a generic scoring scheme for *e-rater* (that excluded the prompt specific dimension of *Topical analysis*) could be successfully used for different prompts. Such a generic scoring method, if successful, could mean that CS could be used to assess NAEP essays in a way that would be linked explicitly to the characteristics of good writing. The set of dimensions shown in Table 2 entitled the *Commonly cited characteristics of good writing* were used by Bennett and Ben-Simon as a representation of 'theoretical meaningfulness' in writing assessment to compare with the *e-rater* features in Table 1.

Table 1 Writing Dimensions and Features in *e-rater* v2

Dimension	Feature
Grammar, usage. mechanics. & style	1. Ratio of grammar errors to the total number of words 2. Ratio of mechanics errors to the total number of words 3. Ratio of usage errors to the total number of words 4. Ratio of style errors (repetitious words, passive sentences, very long sentences, very short sentences) to the total number of words
Organization & development	5. The number of discourse units detected in the essay (i.e. background, thesis, main ideas, supporting ideas. conclusion) 6. The average length of each element as a proportion of total number of words in the essay
Topical analysis	7. Similarity of the essay's content to other previously scored essays in the top score category 8. The score category containing essays whose words are most similar to the target essay
Word complexity	9. Word repetition (ratio of different content words to total number of words) 10. Vocabulary difficulty (based on word frequency) 11. Average word length
Essay length	12. Total number of words

Note. Derived from Attali and Burstein (2005).

Table 2 Commonly Cited Characteristics of Good Writing

Content	Rhetorical Structure/ Organization	Style	Vocabulary	Syntax & Grammar/ Mechanics
Relevance Richness of ideas Originality Quality of argumentation	Paragraphing Coherence Cohesion Focus	Clarity Fluency	Richness Register Accuracy Appropriateness to written language	Sentence complexity Syntactical accuracy Grammatical accuracy Spelling

Note. Derived from Connor (1990) and Johnson. Penny. and Gordon (1999).

Table 3 Initial Mean Dimension Weights Assigned by Members of Committee 1 and 2, along with *e-rater*-H Dimension Weights

Dimension	**Essay 1 Informative**			**Essay 2 Persuasive**		
	Comm. 1	Comm. 2	*e-rater* H	Comm. 1	Comm. 2	*e-rater* H
Grammar, usage, mechanics. & style	13	16	43	15	15	39
Organization& development	37	36	14	37	38	9
Topical analysis	28	35	6	26	33	12
Word complexity	11	9	8	11	9	10
Essay length	11	4	30	11	5	30

The key theoretical aspect of the Bennett and Ben-Simon study was the convening of two separate committees of writing experts who were asked to determine the most appropriate percentage weighting of the *e-rater* v2 dimensions for constructing an overall score. Table 3 shows the weightings offered by the different committees for two kinds of essays (informative and persuasive), and the weightings used by what was called the 'hybrid' use of *e-rater* (*e-rater* H). The hybrid approach used the *e-rater* v2 dimensions so that some 'substantively counterintuitive weights' produced by 'brute empiricism' were controlled or removed. The hybrid approach, for instance, fixed the *Essay length* dimension at 30% so as to not 'overemphasize the influence of this feature on score computation'. According to Bennett and Ben-Simon, the hybrid approach coupled 'a more theoretical motivated feature set with the empirical derivation of weights'. The hybrid approach attained a certain alignment between the expert view of good writing and the optimal statistical prediction of human score by 'brute empiricism'.

Results of the Bennett and Ben-Simon study

Table 3 shows that the two separate committees independently produced very similar percentage weights for the *e-rater* dimensions. The experts gave most weight to *Organization and development* (36% to 38%), followed by *Topical analysis* (26% to 35%). For the writing experts the overall content issues were given about two thirds of the weighting.

The hybrid model gave the highest weight to *Grammar, usage, mechanics. & style* (39% and 43%) and *Essay length* (a fixed 30%). The experts gave 13-16% for *Grammar, usage, mechanics & style* and 4-11% for *Essay length*. It was only in the comparatively low weighting of *Word complexity* that the experts and *e-rater* gave similar weightings (8% and 11%).

Table 3 shows a marked difference between the comparative importance of different aspects of writing quality for writing experts and the operation of a controlled, hybrid weighting of *e-rater* dimensions. Presumably the weightings of 'brute empiricism' (if they had been reported) would have contrasted even more with the writing experts than the hybrid approach.

These results show that even a constrained use of *e-rater* features contrasts markedly with the views of experts about what makes for good writing, and the appropriate analytical scoring of writing. The results confirm the fears of some about the limited validity of CSW in that experts think the quality of writing should be assessed quite differently from the way *e-rater* assesses it to predict human scores (Ericsson, 2006). The results prompt questions about the relationship between the views of writing experts and the scores of humans that can be readily mimicked by *e-rater*. Why is it that the weights given to different features of writing by experts is negatively correlated with the weights used by the constrained model of *e-rater* to predict human scores?

The differences between human and computer scores of writing

The data produced by Bennett and Ben-Simon prompts some analysis of the differences between human and computer scores. A comparison of Tables 1 and 3 shows clear differences between the *Characteristics of good writing* and the *e-rater* dimensions and the features from which they are constructed. The *Characteristics of good writing* give most emphasis to global content issues, and the *e-rater* dimensions are focussed on specific word and sentence level features.

The *e-rater Grammar* dimension matches the dimension of *Syntax* in the *Characteristics of good writing*. It might be claimed that *richness, register, accuracy and appropriateness to written language* of *Vocabulary* in the *Characteristics of good writing* would seem to be roughly reflected in the *word repetition, familiarity of vocab and average word length* of the *e-rater* features. The controlled *e-rater* feature of *Essay length* is not mentioned as a *Characteristic of good writing.* As with the reactions of the experts presented in Table 3, the *Characteristics of good writing* are much more focused on global content issues than the *e-rater* dimensions and features.

The extent to which writing ability is a matter of global content issues (as distinct from more specific language issues) is the crucial matter in arguments about the validity of writing assessment, and hence the validity of computer scoring. Some conceptions of writing ability put global content issues at the heart of the construct, but other conceptions place most emphasis on language as the heart of writing ability. This crucial difference in emphasis is clear in the Bennett and Ben-Simon study, and the substance and significance of it should be recognised and understood. The nature and operation of CSW has brought this crucial difference in emphasis about what is meant by the term 'writing ability' into high relief.

The Bennett and Ben-Simon study prompts the question whether the human marking that *e-rater* is said to predict with 'between 87% and 94% agreement' is more like the *e-rater* dimensions than the views of writing experts or the *Characteristics of good writing*? The ability of *e-rater* to predict some human scoring may depend on the way the humans score the texts in the first place.

The content/thought versus form/language dualism in writing assessment

The results of the Bennett and Ben-Simon study presented in Table 3 show the key dualism of writing assessment. The writing experts give most emphasis to the global content issues of *Organisation and development* and *Topical analysis,* and *e-rater* gives most emphasis to specific language issues of *Grammar, usage, mechanics and style.*

Of the features making up the *e-rater* dimensions shown in Table 1 *Grammar* and *Word complexity* seem fairly directly aligned with the dimension labels, but the features offered to account for *Topics analysis* and *Organisation and development* are very indirect. Features 7 and 8 for *Topic analysis* are based on what is described as 'the bag of words' approach. *E-rater* assesses *Topical analysis* with two kinds of 'content vector analysis' that compare the vocabulary in an essay with the vocabulary in a group of training scripts scored by humans.

There is reason for asking why it can be assumed an assessment of the quality of *Topical analysis* offered by a writer is well represented by the extent to which the words used by that writer converge with the words used by other writers on the same topic who score highly, or vice versa. Even if such topical analysis is good predictor of overall performance (note that it is given the lowest and second lowest weighting by the *e-rater* hybrid), it sets up a dangerous backwash.

Topic analysis is not very important in the *e-rater* hybrid (it is much less important than *Grammar* and *Length*), but if a candidate is to maximise her or his performance on *Topic analysis* when scored by *e-rater* they should set out to envisage and use the kind of words they think better candidates will use in analysing the topic. This would presumably mean that a candidate should try to use words that have a certain complexity - that is they should set out to use long words that are rarely used. As Perelman has pointed out, such an aim contrasts markedly with the commonly cited views of George Orwell about the characteristics to be aimed for in writing well (Jaschik 2011). Aiming to maximise one's score with *e-rater* would encourage the opposite of the highest aim of a writer, which is to say something complicated as simply as possible. The great danger of CSW is that it takes linguistic complexity for complexity of thought, and encourages the dangerous view that sophisticated thought must lead to or be expressed in complicated writing.

Complicated language is not, of course, the same as sophisticated thought. *E-rater* may in effect mean that 'improving' a piece of writing is no more that a process of linguistic inflation.

Topical analysis and *Word complexity* are not weighted very highly by *e-rater*. The actual weightings of hybrid *e-rater* would encourage test candidates to write as much as possible while making as few errors in grammar, mechanics and usage as possible. *E-rater* is clearly focused on errors in sentence level language use and the amount written.

The other global content issue in the *e-rater* dimensions is *Organization and development*. This label is given the most weighting by the writing experts, and the lowest and second lowest weighting by *e-rater*. The features that are used to construct *Organization and development* for *e-rater* are arguably more problematic than the 'bag of words' approach used in *Topical analysis*. The aim of the *Organisation and development* dimension of *e-rater* is to discern a definite structure in the text of 'background, thesis, main ideas, supporting ideas, conclusion'. This dimension is concerning (even alarming?) because it seems to be based on a simple and mechanical view of text organisation. Attempts to explain how this analysis of organisation and development takes place seem to reify the reductive notion of the 'five paragraph essay form' and the topic sentence approach to paragraphing. Fortunately this kind of analysis is not very important in the *e-rater* weighting (it isn't very good at distinguishing between stronger and weaker students), but it does seem to say to candidates (and those who coach them) that a paragraph and an essay are expected to have a certain kind of structure.

What do we mean by writing ability?

The research of Bennett and Ben-Simon into the operation of *e-rater* dimensions highlights the need for clear definition of what is meant by the term writing ability. The term writing ability and the term literacy are often used in vague and hence problematic ways. It is arguable that the *Commonly Cited Characteristics of Good Writing* in Table 2 can be presumed to consider the five characteristics as more or less equally important. The characteristics of *Content* and *Rhetorical structure/organisation* are global content features of text, and the *Vocabulary* and *Syntax and Grammar/mechanics* are specific language features. *Style* seems to bridge the fundamental dualism in that *Clarity* and *Fluency* can be seen as characteristics of both global thought and language use.

How the writing experts would have weighted the *Commonly Cited Characteristics of Good Writing* is an interesting question that it seems was not addressed by the Bennett and Ben-Simon study. Rather than treating them as equally important, it seems reasonable to infer that they would have weighted *Vocabulary* and *Syntax and grammar* as the lowest, and they would have given comparatively high weightings to *Content* and *Rhetorical structure*. What they would have taken *Style* to mean and to include is uncertain. What we do know from the Bennett and Ben-Simon study is that the writing experts would weight *Content* and *Rhetorical structure* highly, in contrast with *e-rater*.

Bennett and Ben-Simon do not discuss the issue of what is meant by writing ability as such, but their exploration of the issue of the weighting of dimension in constructing an overall score from *e-rater* is exploring the issue of construct definition. What their study shows is that writing experts give most emphasis to global content issues over specific language issues.

It could be argued that the proposed dualism of content/thought versus form/language is a false dichotomy because content/thought and form/language are indivisible. While one can see grounds for making such a monist argument (and it probably holds true for the majority of test candidates), it should be recognised that its logical concomitant is global, holistic scoring rather than a process of analytically scoring of separate dimensions or traits. Holistic scoring differs from analytical scoring in that it does not attempt to fix weightings for different dimensions. The way analytical scoring uses a fixed weighting that is applied to each case can be problematic. It is an advantage of holistic scoring that it can use different dimensions to loosely define the construct, but it does not presume that the same fixed weighting of dimensions is applied to all texts. Holistic scoring lends itself to case by case judgements about quality rather than applying the same fixed weighting to all texts. In terms of the Bennett and Ben-Simon study, the idea of weighting of features presumes an analytical process consistent with a dualistic or multi-dimensional view of writing ability. The dualistic view of writing ability proposed here is the simplest and most parsimonious way of recognising the multi-dimensionality of writing ability. This dualism seems to underpin the views of the two expert committees convened by Bennett and Ben-Simon.

The difference between the working definition offered by the writing experts and *e-rater* of writing ability prompts questions about the human scoring that *e-rater* is said to predict as well as humans predict each others. The reported results of the agreement between human markers and *e-rater* are largely based on high stakes ETS testing programs. The pressure at work in such assessments may well have an impact on what is actually assessed in them, and the ability of *e-rater* to predict the scores produced by human markers in those conditions. By comparison, it is worth recalling that *e-rater* did not predict human scores as well as humans predicted each other in the no stakes NAEP study (Sandene, 2005).

How writing ability is conceptualised and operationalised can differ significantly. Human scoring is not a homogeneous process. As we have seen above, there is reason for thinking that marking regimes may differ significantly in terms of the comparative emphasis they give to global content issues and specific language issues. Content issues in the assessment of writing are difficult to define and specify, while language control issues (as is shown by computer scoring) seem to be more concrete and definite characteristics to assess. It can be the case that markers give an effective priority to language control issues in writing assessment (whatever the scoring scheme implies) because they feel more confident in identifying such specific issues. Some markers can be significantly influenced by errors in language use, and some marking regimes explicitly attempt to ameliorate such overreactions.

It may be the case that the high agreement reported with human markers for *e-rater* is symptomatic of an emphasis on specific language issues rather than global content in that human marking. In high stakes testing where there is pressure to mark for convergence (the performance of markers may be reduced to a matter of how many discrepancies with other markers they generate), this pressure can lead to a mental set in which markers are trying to pick the lowest common denominator amongst other markers. In such a system it might be wise for a marker (if they want a low discrepancy rate) to drift towards emphasising specific language issues rather than the more difficult issues of global content. The drift from content to language is significant problem in briefing markers and managing the marking of writing ability.

Whether the agreement of ETS markers with *e-rater* is conditioned by the culture of the markers or not, Bennett and Ben-Simon showed there is a clear and systematic difference between the way writing experts see good writing (and the way they would have markers reward the different characteristics of writing) and the way *e-rater* determines the quality of writing.

Exploring writing ability and the validity of *e-rater*

There were a number of studies of the validity of *e-rater* scoring in the last decade. The desire for a theoretical base for the assessments undertaken by ETS was manifested in a long-term research and development project called the Cognitively Based Assessment of, for, and as Learning (CBAL). As part of the CBAL project Deane et al. undertook a review of the literature on writing cognition, writing instruction, and writing assessment with the goal of developing a framework and competency model for a new approach to writing assessment (Deane, 2008). Deane identified three strands for conceptualising writing ability.

- Strand I is language and literacy skills for writing
- Strand II is writing-process management skills
- Strand III is critical thinking for writing

According to Deane, the research literature reveals a strong emphasis on ‘writing as an integrated, socially situated skill that cannot be assessed properly without taking into account the fact that most writing tasks involve management of a complex array of skills over the course of a writing project, including language and literacy skills, document-creation and document-management skills, and critical-thinking skills’. The broad notion of writing ability presented by Deane gives as much emphasis to critical thinking and social cognition as to literacy as knowledge and use of language. Deane suggests that writing ability is as much a matter of thinking skills as it is a matter of language skills. According to Deane, this view of writing ability is at odds with testing of ‘relatively simple, on-demand writing tasks’.

Deane offered a model for the assessment of writing based on ‘modern cognitive understandings; built around integrated, foundational, constructed-response tasks that are equally useful for assessment and for instruction; and structured to allow multiple measurements over the course of the school year’. This work by Deane produced the kind of definition of domain proficiency Bennett had envisaged as necessary for the grounding of CSW.

The construct coverage of *e-rater*

In the process of evaluating the ‘construct coverage’ of *e-rater*, Quinlan et al. claimed that the *e-rater* dimensions reflected the scoring rubrics of the GRE and TOEFL (Quinlan, 2009), and the 6-trait scoring model of Spandel & Stiggins (Spandel, 1990).

- Ideas and content
- Organization
- Voice
- Word choice
- Sentence fluency
- Conventions

Having reviewed the 6-trait model for analytical scoring and the holistic scoring rubrics of various assessments of persuasive writing, Quinlan et al. concluded that:

> *Typically, one or two traits specify high-level concerns, such as the quality and organization of ideas, while two or three other traits specify low-level issues, such as sentence fluency, word choice, and conventions. The appearance of these general traits, time and again, suggest that there is much consensus about the definition of essay*

> *quality. The definitions may differ slightly across contexts, but the underlying traits appear relatively stable.*

Quinlan reviewed conceptions of writing as 'high level problem solving' and concluded that:

> *Although the e-rater engine may not be able to distinguish between the problem solving of more- and less-skilled writers, it may measure aspects of basic writing skill. Although a written text does not provide a perfect picture of the writer's thinking, it does reveal something about the writer's ability to compose grammatical, well-punctuated sentences - the* sine qua non *of being a skillful writer.*

This view seems to offer a kind of hierarchical view of writing ability. There are basic writing skills that are the '*sine qua non*', or essential pre-condition for skilful writing. The notion that writing ability is some kind of hierarchy of skills is open to question. Whether one sees writing ability as a hierarchy of skills depends on what one is prepared to recognise as skilful writing. It may be the case that those judged to perform well on a writing test show skill in thinking and in expressing their thoughts, but for some candidates there are significant differences between the quality of their thinking and their ability of express their thoughts. The 'ability to compose grammatical, well-punctuated sentences' is not a pre-condition for expressing sophisticated ideas in writing. If such specific language issues were a pre-condition for expressing sophisticated thinking in writing, assessing writing would involve fewer trade-offs and would be much easier that it actually is.

A hierarchical view of the assessment of a piece of writing would begin by asking: 'Is this an correct and fluent piece of writing?' If the answer was that it is not correct and fluent, a low grade would be awarded. If the answer was: 'Yes it is a correct and fluent piece of writing', then it would be asked whether the piece is expressing sophisticated ideas. If the answer was that the ideas are not sophisticated, then the score would be middling. If the answer was that the ideas are sophisticated, the text would be given a high mark. This little parody is not, of course, the way an assessment of writing ability can or should be done. What the parody shows is that what might be called 'basic writing skills' are not some kind of essential pre-condition for the higher levels of skilful writing.

After reviewing the factor analysis of TOEFL by Attali and Powers (Y. Attali, & Powers, D. E. , 2008), Quinlan et al. concluded that '*e-rater* features capture low-level aspects of essay quality, such as sentence complexity, vocabulary, and conventions'. Quinlan et al, concluded that:

> *Whether defined in terms of process or product, the e-rater scoring engine provides partial coverage of the construct, with the majority of measurement capturing the low-level aspects of essay quality that reflect basic writing skills. Future development should address suspected accuracy issues, then turn toward deepening and extending the coverage of traits of essay quality (e.g., 6-trait scoring).*

This conclusion is a clear recognition of the limitations of *e-rater* scoring. As suggested above, there is reason for questioning the claim that e-rater at least assesses the *sine qua non* of writing ability.

Writing assessment and cognition

As an extension of his earlier work Deane proposed in 2011 ‘a socio-cognitive framework for connecting writing pedagogy and writing assessment with modern social and cognitive theories of writing’ (P. Deane, 2011). According to Deane, this general framework:

> *… highlights the connections between writing competency and other literacy skills; identifies key connections between literacy instruction, writing assessment, and activity and genre theories; and presents a specific proposal about how writing assessment can be organized to promote best practices in writing instruction.*

Deane presented reading, writing, and critical thinking as ‘different but complementary activity types that share a common underlying skill set’:

> *It would be possible simply to equate reading with receptive skills, writing with expressive skills, and critical thinking with reflective skills, but that would be an oversimplification. … In other words, reading, writing, and critical thinking appear to be mutually supporting and highly entangled.*

The model outlined by Deane presents reading, writing, and critical thinking as ‘distinct activity systems founded upon common underlying skills’:

> *One can have critical thinking without reading or writing (for there is no requirement that reflective thought be expressed in written form). Writing can take place without deep reflection, for there is no guarantee that the thoughts expressed in a written text will be significant, relevant, fair, clear, precise, complex, accurate, or logical. Yet the whole point of skilled writing is to mobilize all of the resources available to the writer to achieve meaningful goals. The expert writer knows when to apply reflective thinking to writing tasks, just as the expert thinker knows when to use writing as a tool for reflection. The skills are not the same, but they mobilize similar underlying abilities.*

The interpenetration and interdependence of reading, writing and thinking leads Deane to the ‘paradoxical conclusion that a writing test ought to test more than writing’. Deane discussed writing tasks that will focus on ‘content’ and will be ‘driven by critical thinking and rhetorical requirements, and not by surface form’. For assessing these tasks Deane envisages a two part analytical scoring focussed on critical thinking and language. He envisages such scoring would offer informative and useful feedback:

> *It may be particularly instructionally useful for teachers to be able to identify students who are not following the usual trend where fluency and strategic thinking develop in close synchronization. These may reflect special cases, such as students with high verbal abilities in another language or students who need to be challenged to go beyond fluency to engage writing at a deeper level ….*

The proposed dual scoring scheme would ‘encourage instruction that recognizes the importance of developing fluent text production while teaching appropriate writing and thinking strategies’. Deane claims that is possible to make ‘fairly clean separation’ between ‘rhetorical purpose and strategic thinking’ in writing and ‘fluency in text production’.

With this separation in mind Deane considers the prospects of CSW in the following terms:

> *The availability of automated scoring raises the possibility of scoring writing assessments quickly and providing very timely feedback as well as making periodic assessment more feasible and affordable. It also has the potential to distort writing instruction in favor of teaching students those features of writing most easily scored by machine. However, automated scoring may be able to support revision strategies and provide automated scoring of student use of the conventions of written standard English. With automated scoring handling this aspect of text structure in the background, it will be easier to use human scoring to foreground writing processes and critical thinking for writing, skills for which human scoring is most likely to be instructionally useful.*

Deane envisages making effective use of computer essay scoring techniques without substituting computed scores for human judgment about content and critical thinking. Such a system would use computer scoring for fluency-related constructs, which could free human scorers to focus on rhetorical success, conceptual content, and other features that cannot be analysed well, if at all, by computers.

In 2011 Deane also gave further consideration to the use of CSW in a paper entitled *Automated Scoring Within a Developmental, Cognitive Model of Writing Proficiency* (P. Deane, Quinlan, T., & Kostin, I, 2011). He noted that CSW has been criticised for 'focusing attention too much on mechanical correction of errors rather than encouraging critical engagement with content. In other words, it is no simple matter to design summative assessments that are truly tests worth teaching to'.

He discussed a proposal for 'foundational task sets' that are intended to 'instantiate best practices in writing instruction, with the sequence of tasks that may appear on a summative assessment particularly focusing on prewriting, inquiry activities, writing for content learning, and effective use of writing strategies, while also providing appropriate rubrics and models'. He outlines the results of a study by Attali and Powers (Y. Attali, & Powers, D. E. , 2008) that tried to 'to ground *e-rater* in empirical developmental data, and use it to build a developmental writing scale applicable to 4th through 12th grades'. The key findings of Attali and Powers were that it was possible to develop a single writing scale that uses *e-rater* features to score student essays from 4th to 12th grades. The analysis of such material revealed three underlying factors:

- fluency (as measured by essay length and the *e-rater* style feature);
- conventions (as measured by the *e-rater* grammar, usage, and mechanics features); and
- word choice (as measured by the *e-rater* vocabulary and word length features).

As Attali and Powers noted, the scale is 'based only on the specific features used in this study, and thus is limited by what those features measure. Even though past research showed very high correlations between *e-rater* scores and human scores, it is clear that important dimensions of writing are not represented in this feature set'.

On the basis of this research Deane concluded that the *e-rater* features appear to measure Strands I (language and literacy) and II (writing-process management skills) of the three strand competency model of proficiency in writing. The factor analysis showed that the '*e-rater* organization and development features measure aspects of Strand II whereas the grammar and mechanics features measure aspects of Strand I in the CBAL competency model'. He goes on to note that 'no similar pattern exists with respect to the Strand III scores assessing critical thinking for writing', and concludes that:

> *These results are consistent with the hypothesis that none of the* e-rater *features directly measures critical thinking for writing, and they predict Strand III scores indirectly through Strand I and II correlates.*

Deane takes these results as suggesting a 'scoring model in which human and machine scoring served different functions':

> *The e-rater score would focus on text production skills, and a cross-checking human score would focus on critical thinking skills. If larger studies confirm a high correlation between the two, the overall essay score could be presented as a composite of the human and machine scores, with a second rater called in only where the two disagreed, to confirm that the difference in scores actually reflected a difference in the quality of text features versus quality of the underlying content. At this point, this possibility must be viewed largely as a speculation into fruitful future directions, and many complications would have to be considered and resolved before such an approach could be adopted. It would, however, provide one way to leverage automated scoring while preserving the commitment that the CBAL approach to writing has made to the importance of rhetorical, content, and critical-thinking based elements to the writing construct and hence to writing instruction.*

Dean canvasses the following use for computer and human scores:

> *In the long run it might be best to use automated scoring to deal with some parts of the writing construct for which automated analysis is most effective and to reserve human scoring for those aspects of the writing construct that cannot directly be measured by automated means.*

Monaghan & Bridgeman (Monaghan, 2005) had discussed the use of *e-rater* as 'a means of quality-checking human scores', and Enright had described *e-rater* scores as 'complementing human judgment' (Enright, 2010). The proposal of Deane differs from these uses of *e-rater* in that it is based on a recognition of the limitations of *e-rater* scoring (and the scoring of humans that *e-rater* can readily predict?), and it is proposed to focus the humans and the computer on different aspects of writing. Deane writes that these ideas will have to be explored and trialled in different ways to see how they work and what the different scores can be used for. Clearly such trailing is a long way from the 'brute empiricism' that is only concerned with the ability of CS to predict global human scores.

The open questions

This selective overview shows that the way ETS researchers have thought about writing ability and *e-rater* is informative, and much to the credit of ETS. Bennett has recently published certain reflections about the need for non-profit institutions like ETS to recognise that their tax exempt status means they have social responsibilities. The story outlined above shows a degree of responsibility about the use to CSW at ETS that contrasts markedly with the purely commercial purveyor of CS services.

The work of ETS researchers on CSW has reflected an increasingly sophisticated view of writing ability. The initial black box version of *e-rater* was made potentially transparent in *e-rater* v2. The attempt of find a 'theoretically meaningful' grounding for *e-rater* has shown that the software can deal with the specific features of language, and do so in a way that is possibly more accurate and useful than humans can do. But questions about the part that specific features of language should have in an assessment of writing ability remain open. Clearly the issue of the importance of specific features of language in writing ability has to be constantly considered in all marking of writing, if not definitely determined as a fixed weighting. It may be the case that there should be no fixed weighting for content/thought versus form/language. It may be best to recognise that for humans there need be no fixed weighting for different dimensions in assessing the quality of writing, and that human marking is a matter of case by case judgements rather than the application of an algorithm, in the manner of a computer.

The specific and circumscribed use of *e-rater* proposed by Deane raises questions about the current uses of *e-rater,* and other CS software, in high stakes testing programs. What does the evident difference between the way *e-rater* scores texts and how writing experts think texts should be scored shown in the study by Bennett and Ben-Simon suggest about the current uses of *e-rater* in testing programs?

It might be said that what is shown about *e-rater* in the study of Bennett and Ben-Simon, and the conceptual work of Deane, does not apply to the use of *e-rater* to predict holistic scores given by humans in high stakes testing programs. On the other hand it might be said that the fact that *e-rater* can predict the human scoring of high stakes tests might lead one to look more closely at the way those high stakes tests are scored by humans. To what extent is the scoring of those high stakes test focussed on specific language issues? And to what extent does the scoring assess the global thought expressed in writing? Does the assessment of thought and language in the high stakes tests scores that are predicted with 'between 87% and 94% agreement' by *e-rater* give tacit priority to the assessment of specific language issues? Could it be the case that the less human markers agree with *e-rater* the richer and broader is the human marking?

Such questions lead back, of course, to the issue of what is meant by the term writing ability. According to the analysis of Deane, *e-rater* can only directly address two of the three strands of writing ability, and *E-rater* cannot deal directly with the fundamental aspect of critical thinking in writing.

It would be interesting to see what it will mean to focus the human assessment of writing done by ETS on the critical thinking strand of writing ability as envisaged by Deane. (Some marking regimes currently set out to focus on the quality of thinking in test scripts, as in the AST test mentioned above, and give subsidiary status to specific language issues.) How would the marking of high stakes writing tests at ETS have to change to focus on critical thinking? What kind of adjustment would it take to get human markers to focus on critical thinking, and elide language issues? How much adjustment would it take for markers to focus on assessing critical thinking in writing? How would such a change in human marking impact on the ability of *e-rater* to predict human scores?

Note 1: There is an interesting division in the overtones of the terminology. Supporters refer to 'automated essay scoring' and critics refer to 'machine scoring'. On mature reflection 'computer scoring' will be used here as the appropriately neutral term.

References

Attali, Y., & Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. *The Journal of Technology, Learning, and Assessment*.

Attali, Y., & Powers, D. E. (2008). *A developmental writing scale*: Princeton, NJ: ETS.

Ben-Simon, A., & Bennett, R.E. (2007). Toward theoretically meaningful automated essay scoring. *Journal of Technology, Learning, and Assessment, 6(1).*

Bennett, R. (2004). *Moving the Field Forward: Some Thoughts on Validity and Automated Scoring*. Princeton: ETS.

Bennett, R., & Bejar, I. (1998). Validity and Automated Scoring: It's Not Only the Scoring. *Educational Measurement: Issues and Practice, 17*.

Bennett, R. E., & Ben-Simon, A.. (2005). *Toward theoretically meaningful automated essay scoring*: Educational Testing Service.

Burstein, J., Kukich, K., Wolff, S., Lu, C, & Chodorow. M. (1998). *Computer Analysis of Essays*. Princeton NJ: Educational Testing Service,.

Chung, G. K. W., & Baker, E. L. (2003). Issues in the Reliability and Validity of Automated Scoring of Constructed Responses. In M. D. Shemis, & Burstein, J. C. (Ed.), *Automated Essay Scoring: A Cross-disciplinary Perspective*. N. J.: Lawrence Erlbaum Associates.

Deane, P. (2011). *Writing Assessment and Cognition*. Princeton: ETS.

Deane, P., Quinlan, T., & Kostin, I. (2011). *Automated Scoring Within a Developmental, Cognitive Model of Writing Proficiency*: ETS.

Deane, P., Quinlan, T., Odendahl, N., Welsh, C., & Bivens-Tatum, J. (2008). *Cognitive models of writing: Writing proficiency as a complex integrated skill. CBAL literature review writing* Princeton, NJ: ETS.: ETS Research Report No. RR-08-55.

Edelblut, P., . & Mikulas, C. (2005). *A Comparison of the Accuracy of Bayesian and IntelliMetric™ Automated Essay Scoring Methods to Score Essays Written in English.* Paper presented at the 10th Annual National Roundtable Conference, Melbourne, Australia.

Enright, M. K., Quinlan T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring *Language Testing, 27*, 317-334.

Ericsson, P. F., ,Haswell, R. (2006). *Machine scoring of student essays: Truth and consequences*. Logan, UT: Utah State University Press.

Horkay, N., Bennett, R., Allen, N., & Kaplan, B. (2005). Online assessment in writing. In B. Sandene, Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (Ed.), *Online assessment in mathematics and writing: Reports from the NAEP Technology-Based Assessment Project*. Washington, DC: NCES.

Jaschik , J. (2011). Can You Trust Automated Grading? *Inside Higher Ed*. Retrieved from http://www.insidehighered.com/news/2011/02/21/debate_over_reliability_of_automated_essay_grading

Keith, T. Z. (2003). Validity of Automated Essay Scoring Systems. In M. D. B. Shermis, J. C. (Ed.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Mahwah, NJ: Erlbaum.

McCurry, D. (2010). Can Machine Scoring Deal with Broad and Open Writing Tests as Well as Human Readers? *Assessing Writing, Volume 15*(Issue 2)

Monaghan, W., & Bridgeman B. (2005). E-rater as a Quality Control on Human Scores. from http://www.ets.org/Media/Research/pdf/RD_Connections2.pdf

Powers, D. E., Burstein, J. C., Chodorow. M. & Kukich, K. (2001). Stumping E-Rater: Challenging the Validity of Automated Essay Scoring. ETS Research Report 01-03. from http://www.ets.org/research/dload/powers_0103.pdf

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct coverage of the e-rater scoring engine*. Princeton, NJ: ETS. .

Rudner, L. M., Garcia, V., Welch, C. (2005). *An Evaluation of IntelliMetric Essay Scoring System Using Responses to GMAT AWA Prompts*: GMAC Research Reports.

Sandene, B., Horkay, N., Bennet, R. E., Allen, N., Braswell, J., Kaplan, B. & Oranje, A. (2005). *Online Assessment in Mathematics and Writing: Reports from the NAEP Technology-Based Assessment Project*: NAEP.

Shemis, M. D., & Burstein, J. (2003). *Automated Essay Scoring: A Cross-disciplinary Perspective*. N.J.: Lawrence Erlbaum Associates.

Spandel, V., Stiggins, R. J. (1990). *Creating writers: Linking assessment and writing instruction*. NY: Longman.

Vantage. (2004). IntelliMetric™Web-based Essay Scoring Engine. Retrieved from http://www.vantage.com/pdfs/intellimetric.pdf