the jaltcalljournal

ISSN 1832-4215
Vol. 7, No.3 Pages 271–292

An evaluation of automated writing assessment

Craig Hagerman
Osaka Jogakuin College
craig@hagerman.ca

This paper describes research done to empirically study the efficacy of using Criterion, a commercial web-based writing evaluation platform, as a revision aid, especially within a class following the Process Writing pedagogy. This paper first describes the background and rationale for Criterion automated writing evaluation, the Process Writing approach, as well as the natural language processing techniques used by Criterion. In this study an essay was submitted to Criterion multiple times with variations in a single feature each time. By logging the results and comparing how Criterion scored each submission it is possible to determine what components of an essay influence Criterion's scoring. The conclusions from these results suggest that Criterion alone is not an adequate revision aid for the developing writer.

Introduction

The commercial web-based software package, Criterion, is marketed to and used at institutions throughout Japan as a pedagogical aid for second language writing instruction. Despite its widespread use there has been no research to determine how effective Criterion truly is for this purpose. This paper describes research to empirically determine Criterion's suitability as a revision aid.

Criterion introduction

Criterion and *e-rater* were developed for practical reasons – as an automated essay scoring service for high-stakes, time-limited, essay writing. Criterion is an automated essay scoring software package created by Education Testing Services (ETS). It is

Regular Paper

advocated as a useful and effective writing tool for teachers as well as students. Although most of the users of Criterion as native English K-12 students, it is also marketed as a beneficial instrument in second language writing instruction. On the server side, Criterion is powered by a Natural Language Processing (NLP) suite of software (*e-rater*) which is also used by ETS to score the writing portion of the SAT, TOEFL, TOEIC and other exams they offer.

Research question

This research seeks to answer two related questions about Criterion. Firstly, is Criterion effective as a revision aid for second language students and how may it be used most effectively for this purpose? Secondly, does it support and fit well within a Process Writing approach (see White & Arndt, 1991)? Informally, a number of writing teachers have expressed reservations about using a computer system to either score or give revision advice on student compositions. This research seeks to discover to what extent such teacher's misgivings can be supported with evidence, and in what ways Criterion may be used most effectively.

Limitations

Criterion / *e-rater* are used by millions of users for both high stakes and low stakes evaluations, and by both native and L2 writers. This research examines the use of Criterion for low-stakes/ no-stakes writing evaluation of EFL students, more specifically, use of Criterion as a revision tool. Additionally the main concern is with how to improve a medium score, that is, what to tell students so that they can improve a score of 3, 4, or 5 by one or two points. To that end, the ways in which variations in a single essay affected the score and feedback to determine what factors improve the score were examined.

Related work

There has been relatively little research on automated essay evaluation. All of the published research is either authored by computational linguists who work for ETS (a biased perspective perhaps?) or by people from the Humanities involved in writing instruction. There are two books which have looked at the effectiveness of automated writing evaluation software. The first is *Automated Essay Scoring: A Cross-disciplinary Perspective* by Mark D. Shermis and Jill C. Burstein, two NLP researchers employed by ETS. The second is *Machine Scoring of Student Essays: Truth and Consequences* by Patricia Freitag Ericsson and Richard Haswell, which is a collection of articles looking at automated writing evaluation from the perspective of teachers and administrators. The former is a work aimed at psychometricians and computational linguists which contrasts the major automated essay scoring suites on the market and discusses related psychometric issues. It lacks any input or perspective from writing instructors. The latter book contains sixteen articles with reflections and critiques of automated essay scoring software from the perspective of composition instructors. The present research bridges the gap between the two camps by being undertaken by a computer scientist with experience teaching writing to second language learners.

Background

Criterion

History. ETS describes Criterion as "a web-based service that evaluates a student's writing skill and provides instantaneous score reporting and diagnostic feedback" (Attali & Burstein, 2005, p. 7). Criterion is the name of a web service (application) offered by ETS. It is composed of two pieces of software on the backend: *e-rater* and Critique. *e-rater* gives the essay a holistic score, Critique gives feedback on Grammar, Usage, Mechanics, Style and Organization.

E-rater was first developed by ETS in 1999 to score high-stakes essays on the GMAT test. ETS is a large not-for-profit educational testing company that produced many standardized tests, including TOEFL, TOEIC, GMAT, GRE, and SAT. Scoring written essays was a logistical problem for them involving hiring and training thousands of graduate students, to score the essay portion of exams. On the other hand, all of the other portions of those exams could be automatically scored by computers. Thus, essay-scoring software was developed to alleviate the human resources imbalance involved in exam scoring.

ETS added a web-based front-end to *e-rater* and released it as a subscription service under the name Criterion in 2001. By 2002 they had 50,000 users. Currently, they have over a million users worldwide. The English as a second language exams offered by ETS (TOEFL, TOEIC) are very popular in Japan, where about 1.5 million Japanese take the TOEIC test each year, and around 80,000 sit the TOEFL exam. Writing teachers in Japan might assume that Criterion was created for L2 students, but in fact, the vast majority of users are in over 3000 K-12 schools in the USA. The primary market is thus for native-English student writers. Similarly, the research done on Criterion and other automated essay evaluation software has been largely confined to researching the efficacy of the software with native compositions.

Using Criterion. After logging into Criterion, an instructor can select an essay type, title and prompt and control various settings (see Appendix C). For example, the instructor can control whether and how long a time limit is imposed, and what feedback will be provided. When the student finishes (or the time expires) the essay is submitted and a short while later the students receives a holistic score and *Trait Feedback Analysis*. The holistic score is a score out of 6, with 6 being the highest score and 1 being the lowest. An essay which Criterion cannot score receives a holistic score of '**N/A**'. *Trait Feedback Analysis* is also given for five trait categories: grammar, usage, mechanics, style, and organization & development (see Appendix D).

An instructor may choose a Criterion-provided prompt from one of sixteen categories (grade 9, College second year, TOEFL, etc.) corresponding to different levels and writing styles of students. Additionally an instructor may provide their own, original essay topic and prompt.

Discourse structure. Criterion assumes that the essay follows the five-paragraph strategy, and additionally that a well-written essay will contain discourse elements: thesis, main ideas, supporting ideas, and conclusion. It is trained on a large corpus of (human) annotated essays so that it can identify those discourse elements. Additionally, it uses a dictionary of transition terms plus heuristics to identify phrases that start a new discourse segment,

such as "In summary", "First", etc. Recognition of and feedback on discourse elements is a feature requested by K-12 teachers using Criterion to support the scores (Burstein, 2009).

Criterion highlights features a writer might want to revise such as "the use of passive sentences, as well as very long or very short sentences within the essay. Another feature of undesirable style that the system detects is the presence of overly repetitious words" (Attali, 2004, p.2). This is also something requested by K-12 teachers.

Process Writing

Process Writing is the default approach to teaching writing to second language learners, and is the prevalent writing approach in Japanese institutions. It is the approach used in all classes involving writing at Osaka Jogakuin College (OJC), where this research was conducted. At OJC Criterion is used as part of the revision stage of Process Writing. It is likely that this is the case at many other institutions as well since Process Writing is the default writing pedagogy for second language writing instruction. Moreover, ETS claims that Criterion is designed to support the Process Writing pedagogy. As Burstein (2009) writes, "Criterion embodies the process writing approach. This approach supports the idea that students should be able to write several drafts of a piece of writing. Consistent with this, Criterion allowed students to submit multiple revisions of essays, and receive a new score for each revision" (p. 10). Because of the prevalence of Process Writing instruction and ETS's stated support for this pedagogy, this research used empirical data to determine to what extent Criterion actually does foster Process Writing. Therefore, the following paragraphs will explain the salient features of this pedagogy. Stone describes Process Writing as "learning to write by writing" (1995, p. 23). In the traditional approach, writing assignments are a way to practice and reinforce specific grammatical / lexical patterns. The traditional approach focuses on the final product, with accuracy the principal concern. According to Raimes (1983), traditional writing teachers "have trapped our students within the sentence… [and] …respond to the piece of writing as item checkers not as real readers" (p. 167). (Please consider the rewrite as a suggestion.)

Process Writing assumes that anyone can write. That is, the learner does not have to master skills (form) first before writing. Process Writing focuses on the process rather than the end product. Good writers plan, revise, rearrange, delete, re-read, and produce multiple drafts. This is what process writing is all about. In Process Writing the teacher moves away from being a gatekeeper, judge, grader of a finished work. The teacher becomes a reader rather than a marker, responding to content rather than form.

The traditional approach of focusing on language errors does not improve grammatical accuracy or writing fluency (White & Arndt, 1991). What does help improve writing is focusing on the content – on what learners say and the meaning of what they write. Moreover, feedback is more useful to learners between drafts rather than just at the end, during grading of the final product. In Process Writing students should come to see texts they write as changeable and non-static, as a recursive rather than linear activity.

How Criterion works

E-rater and NLP techniques

For the built-in essay prompts, Criterion / *e-rater* has large corpus of human rater (**HR**) scored essays on the same prompt. This means that for each essay prompt there are a large collection of essays which received a score of 6 (best) from human rather, another large collection of 5s, and so on. This is used as a training set by artificial intelligence (software in which the) procedures to create a model of characteristic properties of compositions within each ranking. Then, machine learning algorithms are used to analyze that corpus of essays for relevant features to complete the process. *E-rater* does a stepwise linear regression procedure to find a small set of features that are most predictive of an essay score.

The words used in an essay are highly important. *E-rater* describes an essay in terms of the number of unique words it contains, the average length of words, number of words with 5+ characters, 6+ characters, etc. This kind of data is used to measure an essay's vocabulary in terms of range, frequency, and lexical complexity (set of words / total number of words). It also looks at syllable count and sentence length, which are taken to be a good predictor of readability or text difficulty.

Term frequency-inverse document frequency (tf-idf) is used to give a weight to each word in the training set and this document-wide weighting is then converted into a multi-dimensional vector. It is assumed that good essays on the same topic will frequency contain similarly weighted words, as will weak essays. When a new essay is submitted on one of the built-in prompts, all the same word analysis is done that was done to the training set (word length, number of words, tf-idf, create a vector). The results from the analysis of the new essay are compared to the training set by comparing the cosine similarity of the vectors. The result of this comparison is used to determine the holistic score.

In the case where an instructor writes their own essay prompt there is no training set, since there is no pre-scored corpus of essays with which to compare the new one. In that case **ETS** cannot compare features with an existing model. Instead, *e-rater* uses a general model for comparison and then proceeds in the same way as outlined above (analyzing lexical complexity, tf-idf, etc.).

Bi-grams are pairs of two successive words within a sentence. For example, the sentence "This is a pen" contains the bi-grams "This is", "is a", "a pen." *E-rater* uses bi-grams to find non-grammatical text. The bi-grams of new essay are compared to a large corpus of Standard English using statistical corpus-based models to detect grammatical errors. That is, by comparing a given bi-gram to statistical information on millions of bi-grams taken from English newspapers *e-rater* can calculate how probable or improbable the input is.

Another factor which contributes to the scoring of a composition within Criterion is essay length. The researchers at Criterion acknowledge this saying "it is the single most important objectively calculated variable" (Attali & Burstein, 2005, p. 4). Attali and Burstein contend that there is a strong correlation between quantity and quality in **HR** scored essays. Longer essays are said to usually have more details, support, syntactic variety, and vocabulary range.

E-rater was developed for high-stakes testing assessment with an assumption that the writer should follow the classic five-paragraph essay strategy. This assumption carries over to Criterion's evaluation of a composition. Built-in prompts and default time limits reflect this assumption.

Method

Design

This research sought to address two questions about Criterion. Firstly, how may Criterion be used effectively as a revision aid. Secondly, how well does Criterion truly fit into a Process Writing structured course? In order to answer these questions, we investigated how Criterion works by seeking to identify what composition features contribute to a score of 4, 5 or 6, and discover how manipulating these variables affects an essay's score. In order to discover how Criterion scores, multiple essays were submitted to 3 different prompts multiple times, with each variation involving the manipulation of a single essay feature. The results (holistic and feedback) were noted for each trial and compared across the three prompts. One baseline essay was composed by the researcher and modified to create variants. This is the same procedure used in the only non-commercial quantitative research done on automated writing assessment that we could find (Herrington & Moran, 2001; Jones, 2006; McGee, 2006) and is conducted as an extension of that prior research. Along with the baseline essay, a high-scoring model essay from the **ETS** essay bank was similarly used with variations to provide an **ETS**-originated comparison. Two further off-topic essays were used to show the effect of submitting a composition vastly different in expected content and writing genre.

Essay titles

For this study a 'College 1 Expository' essay prompt was chosen titled "Resources Disappearing" (hereafter Resources). The second essay prompt was a 'College 1 Scored Instructor Topic Expository'. That is, it is an instructor-created prompt at the same level as Resources Disappearing. This essay was titled "Deforestation".

The third essay prompt was also a 'College 1 Scored Instructor Topic Expository" essay, but in this case the title was simply "Check any writing" (hereafter Check). Criterion allows an administrator or instructor to use their own essay title and prompt rather than being limited to the built-in options. At Osaka Jogakuin College all students enrolled in any course with an essay requirement are automatically given a Criterion account for that course. Moreover, in all such courses students are provided with a default prompt within Criterion titled "Check Any Writing". The prompt itself simply says "Copy your essay here to check any writing". This is provided to allow students to get feedback on any writing from another source. Additionally they are encouraged to get feedback from peers, from their instructor, or from the writing center (**OJC** runs a writing center to provide short consultations on any writing questions students might have). Such feedback should then be used to inform subsequent drafts. It is hoped that by providing a generic Criterion prompt students can take responsibility for their own writing by submitting drafts and reflecting on the feedback on their own.

In some courses student make chose from a variety of essay topics, in which case an instructor will often make a single Criterion essay title and prompt for all students. For example, in one module of a course the theme is "Peace studies", and students may chose from ten different essay prompts. In this case, rather than creating ten separate prompts within Criterion (which will create disparate class assessment results) it is common for an instructor to create a prompt titled "Peace" with the prompt "Submit your peace studies

essay here". Informally, it seems that instructors at other universities follow a similar practice.

Essay prompts

The prompt supplied by Criterion for Resources is as follows:

> Many parts of the world are losing important natural resources, such as forests, animals, or clean water. Choose one resource that is disappearing and explain why it needs to be saved. Use specific reasons and examples to support your opinion.

Initially, both Deforestation and Check were given the generic prompt:

> Enter your essay here

However, after an initial trial submitting the same essay to the three prompts it appeared that the prompt itself was affecting the holistic score. Consequently, the prompt for Deforestation was changed to:

> Deforestation is a serious problem today with wide-ranging effects beyond just the cutting down of trees themselves. Write an essay explaining reasons why trees are an essential component of the environment and why deforestation should be stopped.

Model essays

The essay variations submitted to the three prompts were based on two models. One is an example provided by Criterion. After a student submits their essay the instructor can chose to allow them to see example essays for a given prompt to compare their work to another essay that scored a 6, or 5, or 4, etc. Those model essays are presented as an image so that the text cannot be copied and pasted. For a baseline *gold standard* the 6/6 example essay (*Example*) for *Resources* was copied (typed) from such an image. It was felt that an original essay should also be used for comparison since it is possible that Criterion flags submissions of their example essays. The original essay (*Original*) was written by the researcher based on the content and writing style typical of a good L2 student. The *Original* was written on the topic of deforestation (original prompt) to address the problem of deforestation and thereby have a valid basis for comparison with the instructor-created prompt.

Example and *Original* have the sentence and word lengths shown in Table 1:

Table 1: Number of words and sentences in sample and original essays

	# of words	# of sentences	Average # of words per sentence
Sample	517	24	21.5
Original	339	26	13.0

Criterion essay prompts can either be for expository or persuasive compositions. The three prompts as well as the example essay and original essay were all expository. A third essay was used for some of the variations; a persuasive essay on the effects of smoking (hereafter

Smoking). The lexicon of this essay is assumed to be quite different from that in the training sets for *Resources*. In addition, the transitions and other organizational features of such a persuasive essay are distinctly different from an expository essay. Since ETS claims that Criterion analyses lexicon, discourse features and organizational structures it would be enlightening to see how it scored an essay which a human rater would take as different in these three factors.

Essay variations

Over 60 essay variations were submitted to the three prompts. The essay features manipulated in each variation are summarized below. The variables below were selected because they are either features investigated in previous research (Herrington & Moran, 2001), (Jones, 2006), (McGee, 2006) or are features other composition teachers have suggested (informally to the author) that Criterion looks for.

Essay length. The reference essay was made slightly shorter / longer, much shorter / longer or very short / long while retaining the same number of words per sentence. A large number of the essay variants manipulate the essay length variable. Previous research has suggested that essay length was the single largest determinant of an essay's holistic score. Variations on essay length (and sentence length) were submitted to determine if this is (still) true of Criterion, and if so, try to determine what word count Criterion considers optimal.

Sentence length. For each of the overall essay lengths (above), the number of words per sentence was varied, combining sentences to make more words per sentence for the same essay length, or slicing sentences into two or three to make less words per sentence.

Number of sentences. A variation on the above: for a given essay length the total number of sentences was varied while retaining the overall word count.

Paragraphs. The reference essay was submitted with identical sentences, but divided (via a blank line) into 5, 4, 3, 2 or 1 paragraph(s).

Discourse structure. Variations on the reference essay used different topic, thesis and concluding sentences. Each of these was simplified, greatly simplified, weakened (for example, a topic sentence with a focus but no controlling idea) and rewritten to be inappropriate. Each of these was also varied to be longer, or use different transitions or more complex vocabulary. Additionally variations were submitted with each of these elements removed.

Unity & coherence. The reference essay was submitted with paragraphs in reverse order, and with all of the words in reverse order. Other variations involved replacing large sections (one third, half, two thirds or the entire essay) with content from a different essay (Smoking). Other variations involved removing sentences from each paragraph to break the logical flow, or replacing those breaks with sentences from the Smoking essay.

Grammar & mechanics. Essay form was manipulated in a variety of ways since Criterion offers feedback on these elements. Variations introduced errors related to subject-verb

agreement, verb form, word form, preposition errors, article problems, incorrect punctuation, run-on sentences, sentence fragments, incorrect capitalization and altering passive-active voice.

Transitions. The reference essay was altered by removing all transitions, substituting different transitions, using only limited, simplistic transitions, and using a variety of high-level transitions.

Experimental data

The *Example* and *Original* essays were submitted to the three Criterion prompts: *Resources* (Criterion built-in prompt), *Deforestation* (instructor-created prompt), and *Check* (generic instructor-created prompt). The scores are listed in Table 2.

Table 2: Scores for sample and original essays

	# of words	**# of sents**	**words / sent**	*Resources*	*Deforestation*	**Check**
Example	517	24	21.5	6	5	N/A
Original	339	26	13.0	6	4	4 - N/A

The Criterion '6 out of 6' *Resources* sample essay unsurprisingly scored 6 when submitted to the *Resources* prompt. However it was scored lower (5) when submitted to *Deforestation* and was unscorable when submitted to *Check* (**N/A**), with an advisory that it appeared to be on a different topic. The *Original* essay performed similarly, receiving full marks for the Criterion prompt (*Resources*) but dropping down to 4 for both *Deforestation* and *Check*. The essay received a holistic score of 4 from *Check* with an **N/A** advisory again that it appeared to be on a different topic. This was true with all submissions to *Check*. All composition submitted to *Check* received an **N/A** advisory that the essay appeared to be off-topic. Some submissions would receive a holistic score regardless, some not, but in all cases the holistic score was lower than that given to the same composition submitted to *Resources* or *Deforestation*.

The non-expository essay, *Smoking* was used to determine if Criterion could identify an off-topic essay. The scores it received are summarized in Table 3. The Criterion-supplied prompt, *Resources* identified it as off-topic and unable to be scored. However, both *Deforestation* and *Check* scored it without complaint. In fact, this was one of the rare submissions where *Check* did not give an advisory about an off-topic essay.

Table 3: Number of words and sentences in non-expository essay on smoking

	# of words	**# of sents**	**words / sent**	*Resources*	*Deforestation*	**Check**
Smoking	349	20	17.6	N/A	4	4

The holistic scores of the main essay variations are shown in Table 4. Since *Check* only received a holistic score of **N/A** on subsequent variations it is omitted.

Table 4: Holistic scores of main essay variations

	# of words	**# of sents**	**words / sent**	*Resources*	*Deforestation*
no TTC 1	259	20	12.95	5	4
no TTC 2	344	31	11.1	6	4
weak TTC	344	29	11.86	6	4
short 1	285	22	12.95	6	4
short 2	169	15	11.3	5	3
long	532	42	12.66		5
long 2	1047	65	16.1		N/A
long 3	543	38	14.3		5
short sents 1	348	41	8.5	6	4
short sents 2	345	55	6.5	6	4
short sents 3	334	58	5.8	6	4
short sents 4	300	63	4.8	5	3
short sents 5	370	80	4.7	6	4
long sents 1	372	17	22	6	4
long sents 2	405	16	25.4		5
rev sents	339	26	13.0	6	4
rev words	339	26	13.0	N/A	N/A
1 paragraph	339	26	13.0	N/A	N/A
form prob.	322	26	12.5	5	3
1/3 diff.	365	28	13.1	6	4
half diff 1	352	23	15.4	6	4
half diff 2	362	25	14.6		5

Table key

No **TTC** 1 = remove thesis, topic and concluding sentences
No **TTC** 2 = same as **TTC** 1, but add extra words (i.e. word count is unchanged)
weak **TTC** = weak topic, thesis, and concluding sentences
short 1 = slightly shortened (overall word count)
short 2 = very short (overall word count)
long 1 = lengthened (overall word count)
long 2 = very much lengthened (overall word count)
long 3 = lengthened, more transitions, longer sentences, bigger vocab
short sents 1 = shorter sentences
short sents 2 = very much shorter sentences
short sents 3 = very much shorter sentences, no transitions
short sents 4 = very very short sentences
short sents 5 = very very short sentences, but slightly longer word count
long sents 1 = longer sentences
long sents 2 = much longer sentences
rev sents = reversed sentences (reversed order of all sentences)
rev words = reversed words (reversed order of all words and punctuation)
1 paragraph = no paragraph divisions (i.e. one big paragraph)

form probs = introduce grammar / mechanics / usage errors
1/3 diff = one third different content (taken from *Smoking*)
half diff 1 = second half different content (taken from *Smoking*)
half diff 2 = first half different content (taken from *Smoking*)

Results

This study revealed several interesting results which should be of immediate use to writing teachers as well as course coordinators and administrators. The results do not support the claim by **ETS** researchers that Criterion 'embodies' Process Writing except in the most superficial way. It is clearly better to use Criterion-provided prompts rather than instructor-created ones. Disappointingly, for all of the feedback and analysis that is given for an essay, the single greatest determiner of score is simply the total word count. All other factors play a far lesser role in calculating the score.

Prompts

One result which will have important implications for instructors who create generic prompts is that the wording of the prompt can have a far greater affect on a score than the composition itself. The generic *Check* prompt ("Check any writing here") received an **N/A** advisory in almost all trials and a lower holistic score where it was scored. This result makes it clear the sort of general prompts frequently employed at Osaka Jogakuin College and other institutions is inappropriate for use with Criterion scoring.

Another important finding is that the prompt provided by Criterion (*Resources*) consistently scores an essay higher than the very similar instructor-created *Deforestation* prompt. This is understandable. **ETS** gives e-rater access to an enormous corpus of **HR** scored essays, which are used by machine learning algorithms to identify traits and characteristics that can be used to assess a new submission. This is somewhat analogous to how a writing teacher, after years of reading compositions on the same topic develops good intuitions about what an 'A' paper will be like, compared to a 'B' or 'C'. When an instructor creates their own prompt within Criterion, *e-rater* is hobbled by lack of prompt-specific machine learning and must generalize based on all expository essays. Nonetheless, it is disappointing and somewhat surprising that an instructor-created prompt can be expected to consistently receive a lower (more conservative?) score.

When the instructor-created prompt was replaced with the prompt wording of *Resources*, and all variations re-submitted there was no change in scores. The actual details of the prompt instructions would most certainly be relevant and a factor to consider in scoring by a human, but it does not appear to play much of a role for Criterion. Thus it seems it is necessary to have a specific prompt with sufficient detailed instructions, but the actual meaning of the prompt and instructions is less important. It seems likely that behind the scenes *e-rater* is using **NLP** techniques to extract the content words from the prompt and possibly use those words to create a broad semantic web of expected essay content.

Essay length

The essay variations clearly show that the single most important factor in Criterion scoring is raw word count. Longer essays score higher (up to a point), shorter essays score lower.

The primacy of word count in Criterion has been noted by other researchers. (Jaschik, 2007), (Winerip, 2005). The **ETS** researchers defend this by stating that in human scored essays as well longer essay are usually scored higher, because longer essays usually evidence more supporting sentences, greater use of compound and complex sentences and other higher level discourse elements (Burstein, 2009). For the writing teacher giving advice to students on how to improve their Criterion scores the simple admonition to 'write more' would suffice.

The scores from essays of different lengths show that, as far as Criterion is concerned, between 300–600 words is likely an ideal length. All else being equal, an essay with 285 words scored lower (on *Deforestation*) while a very short essay with 169 words scored lower on both *Resources* and *Deforestation*. This length is typical of what many first and second year students at **OJC** produce for a first draft or for a timed Criterion session. Therefore, encouraging students to produce longer works could very well improve their scores (and likely the overall product). It should be noted that the very short essay (167 words) was rated a 5/6 by *Resources* (3/6 by *Deforestation*). It is surprising that an essay this short is scored so favorably.

Essays over a certain length will not be scored by Criterion. The very long essay (1047 words) received an **N/A** and an advisory that it could not be scored. This reflects *e-rater*'s origins as an assessment tool for timed (30 minutes) essay exams. This is another indication how **OJC**'s use of Criterion to allow students to 'check' *any* writing in inappropriate. If a composition is 300–600 words and the instructor creates a prompt Criterion can score the essay, but after first year most students are expected to produce essays longer than 5 paragraphs. Criterion was not designed for this use and will return such a composition as unable to be scored.

It was thought that the number of words per sentence would likely be an important contributor to Criterion's holistic score. After all, within the *Trait Feedback Analysis*, this Criterion produces such statistics for the writer to reflect on. The *Example* and *Original* essays had 21.5 and 13.0 words per sentence respectively. For comparison, statistics were gathered from the Penn Treebank and Brown corpora. (The Penn corpus contains one million words from the Wall Street Journal, Brown is a general English language corpus containing roughly one million words.) The Penn Treebank contains an average of 25.72 words per sentence, while Brown has an average of 22.0 words per sentence. It was found that sentence length did not have an effect on the holistic score unless the sentences were extremely short. An essay with an average of 4.7 words per sentence was scored lower. On the other hand an essay with 5.8 or more words per sentence received the same score (all other factors being equal). However, even with an essay composed of extremely short sentences, increasing the overall word count will cancel out the lower score.

Scoring

Criterion's scoring on *Resources* does not instill faith in its performance. The *Original* essay was rated 6/6 on the *Resources* prompt. Most variations on this essay continued to be rated 6/6. The only lower scores were for a short essay, or many grammar problems. In those cases the score was only lowered to 5/6. An essay with both of these features also scored 5/6. The essay was altered in many ways: absence of discourse elements such as topic, thesis and concluding sentences, swapping in entire paragraphs from the *Smoking* essay, changing or deleting all transitions, shortening or lengthening sentences; yet none

of these variations produced a different holistic score. Even reversing the entire essay such that the last sentence was first, followed by the penultimate, and so on, had no effect on the score. It is disturbing to think that such elements that a human rater would consider important are not reflected in the comparative Criterion score.

A few variations received an **N/A** and were rated as unable to be scored. Those variations involved reversing all words (thus rendering the essay ungrammatical and incomprehensible), submitting the essay as one single paragraph and submitting and entirely different essay (*Smoking*). With respect to the single paragraph essay, the advisory reported, "Your essay could not be scored because some of its organizational elements could not be identified". **ETS** researchers say 'don't try to fool our systems' since Criterion assumes a good faith effort. This seems a fair assumption for high-stakes and even medium-stakes use. The three variations rated as unable to be scored are unlikely to be submitted by students. Rather they indicate that Criterion expects an essay composed of some paragraphs, and (with respect to the built-in prompts), at least some fraction should be on-topic (likely with respect to the expected semantic web for a given prompt), and syntax should be comprehensible (likely with respect to the statistical likelihood of each bigram).

Unity, coherence and discourse elements

Except in the case where a persuasive essay (*Smoking*) was submitted the Criterion-supplied prompt (*Resources*) scored the *Original* variation identically. When the entire *Smoking* essay was submitted, *Resources* returned an **N/A** with off-topic advisory. When the same variations were submitted to the instructor-created *Deforestation* the score was unchanged except in the case where the first half of the essay was substituted for the first half of *Smoking*. In that case, the holistic score actually increased (4 to 5). However when the second half of *Original* was replaced with *Smoking* the same increase was not seen.

It was predicted that Criterion would flag the presence or absence of important discourse elements such as topic, thesis, concluding sentences as well as transitional phrases. However, the results do not support this conjecture. When the topic thesis and concluding sentences were removed, the holistic score did drop, but when (supporting) sentences were added to make the total word count comparable the score returned to the initial value. Similarly, adding or removing transitional phrases did not seem to have an impact on the score.

Trait feedback analysis

Most of *Trait Feedback Analysis* is concerned with form (rather than content). It does identify problems related to grammar, usage, mechanics etc., but also identifies some elements which are not errors. Additionally, some grammar and usage errors were intentionally introduced which Criterion did not catch. Moreover 'Style' flags certain elements for students to consider (such as 'Passive voice') which are not errors, but often seem to confuse students who wonder why a given sentence is (syntactically) 'wrong'. Essay variations based on passive vs. active voice show that there is no difference in holistic score.

'Style' usually flags key topic words in an essay as 'Repetition of words' and encourages the writer to consider synonyms. When talking about deforestation 'tree' and 'forest' are flagged. This is also often a point of confusion for students. If some of those instances are

changed to a synonym (say 'plants' and 'woods') often all four of those words will then be flagged as repetitious. Repetition was found to have no impact on the holistic score.

At first glance, the 'Organization & Development' section of *Trait Feedback Analysis* seems impressive in its ability to correctly highlight the thesis, main ideas (topic sentences), supporting ideas, conclusion and transitional words. Unfortunately the evidence from essay variations shows that there is little real analysis of these elements taking place. Regardless of the content, organization or coherence of an essay, Criterion will almost always automatically flag the thesis, main ideas, supporting ideas and conclusion based solely on the position of sentences within the composition. Thus the last sentence(s) of the first paragraph will be highlighted as the thesis irrespective of whether they are or not. Deleting the thesis, moving it to a different position within the introduction or replacing the thesis with sentences from a different essay will not change the behavior of identifying the last sentence as the thesis. The other elements are identified similarly. This behavior leads the writer to believe that they have a thesis or topic sentence even when a human reader would recognize that they are absent. This is confusing behavior which could hurt more than help the developing writer. Given that this is the sole category of *Trait Feedback Analysis* not concerned with form, it is unfortunate that the implementation does not show any ability to recognize organizational elements on their own. Criterion is much better at identifying transitional words and phrases, but this can be done simply by a look-up dictionary of transitional phrases.

Discussion

The feedback given by Criterion certainly does not support the claim that it embodies the Process Writing approach. The first four categories of *Trait Feedback Analysis* (grammar, usage, mechanics, style) are concerned with form; the main focus of the traditional approach to writing. The fifth category, 'Organization & Development' ought to be of use within the Process Writing approach. The discourse elements of thesis, topic, support and conclusion are important considerations in early Process Writing drafts. Unfortunately, as previously stated, Criterion seems to identify these elements only by means of sentence position within the essay. Giving misleading feedback to the learner is worse than no feedback since she would likely think, given Criterion's highlighting of these discourse elements, that her essay has satisfactory organization and development, regardless of the actual content.

Students frequently submit an essay (to a teacher or Criterion) and then rewrite subsequent drafts based on the feedback and advice given on the prior submission. After submitting an essay to Criterion, students are presented with a holistic score and *Trait Feedback Analysis*. The former has generic comments (i.e. all essays that score 4 have the same holistic comments, etc.). The latter is largely focused on form and insofar as it does focus on content and organization that feedback may well be wrong or misleading. If a teacher does not intervene to provide extra direction the student would naturally be inclined to address the errors and issues raised by the *Trait Feedback Analysis*. This, then leads students to focus their efforts on improving form. (After all, there is no advice given on content issues or how to improve organization.) Furthermore, it is likely a waste of time for a student to spend time correcting punctuation and negations errors, correcting prepositions and verb tenses and reconsidering article use. Improving these elements would likely not improve their holistic score, unless the form was quite weak to begin with. To get a higher score they should simply be given one piece of advice: write more. To truly improve their writing,

students should get feedback from a human, who could note the important content and organizational traits of the essay and give far more valuable advice on how to improve as a writer.

Conclusion

This research sought to answer two questions about Criterion. Firstly, how can Criterion be used most effectively? Secondly, to what extent does Criterion fit in with a Process Writing approach? The most effective use of Criterion is to work with it, not against it. It is not a tool that, like any tool, is most effective within its own domain. Criterion clearly does not use any advanced **NLP** techniques to try to encapsulate meaning of content. Nor is it very competent at assessing organizational features.

The most effective way to use Criterion is to stick with the provided prompts as far as possible. With those built-in prompts student essays can be compared to a large corpus of other essays on the same topic using advanced machine learning algorithms. The data suggests that this will give students a higher holistic score than an identical essay submitted to an instructor-created prompt. Additionally, the built-in prompts are more capable of detecting off-topic essays. Another way to effectively use Criterion is to use its score to encourage learners to write more. Japanese students are frequently slow writers producing shorter text on timed exams. The knowledge that just by writing more their score may increase may well motivate them to expand their composition. Writing instructors know that there are many elements of good writing. Length is only one, and perhaps not a primary one. However, as many writing instructors have discovered, encouraging students to write more often has the side effect of improving their writing overall. Learners who write longer compositions frequently are stronger writers. Therefore, Criterion can be used to push students to write longer compositions, and use longer sentences.

Another way in which Criterion can be effectively utilized is as a pre-final form check. At **OJC** students typically write (some) rough drafts prior to submitting a final paper. In such an environment, in the early stages students could get feedback from peers or their instructor on content and organization and base their revision on those comments. Then, prior to a final submission, Criterion could be used as automatic feedback on form elements. Criterion does have a lot of *Trait Feedback Analysis*. If grammar or transition usage is especially weak, addressing the problems flagged by Criterion could improve the holistic score. This, in turn would make for a stronger essay submitted to their instructor. In this way, Criterion could fit into Process Writing as a late revision feedback tool.

With the exception of the above method of incorporating Criterion into a Process Writing approach, it does not engender this pedagogy at all. Perhaps the **ETS** researchers who made such a claim merely mean that Criterion can be set up to allow students to make multiple submissions (rather than the single submission allowed during high-stakes testing). This, in itself, is not Process Writing. Rather Criterion encourages the learner to focus on the form-based elements which are the hallmark of the traditional approach to writing instruction. Understandably, the limitations of current **NLP** abilities preclude Criterion from rating or giving feedback on the content. The learner is thus left with the impression that errors related to form are where they should focus their revisions.

Criterion suffers serious weaknesses as a revision tool. Criterion is incapable or recognizing and assessing some of the most important features of a well-written composition. There is no recognition of unity and coherence. An essay with reversed is scored the same

non-reversed version. Having large portions of the essay be on a completely different topic and essay genre does not adversely affect the score. (In fact, in one case the score was raised by inserting off-topic paragraphs.) Inappropriate or absent thesis, topic and concluding sentences and transitional phrases is not recognized. These weaknesses are mitigated by using a supplied prompt rather than an instructor prompt. Yet, this only improves the holistic score by allowing comparison with features identified in the training set corpus and does not improve the quality of the *Trait Feedback Analysis*. For its designed purpose of assessing high-stakes tests, Criterion does an adequate job. However, the results of this study show it to be inadequate as a revision aid within a Process Writing pedagogy for second language learners.

Future work

ETS adds features incrementally and does an upgrade each year. They have already added ability to detect and comment on form errors typically made by Japanese students. They have mentioned that they want to develop the ability to comment on how good a thesis statement is, and do better detection of on-topic thesis statements. So it is getting better bit-by-bit.

It is very difficult to write an essay that receives a holistic score of 1. The explanation of Criterion feedback provided by ETS says such scores are rarely seen. Many Japanese university students get scores of 2 or 3. This research was largely confined with trying to identify ways to improve an essay with a score of 4–5. In future work it would be useful to try to identify what qualities of an essay would receive a score of 1 and how to improve lower scores that so many intermediate Japanese students receive.

Another avenue not investigated in this research is to look into whether the level of vocabulary used according to the Academic Word List correlates with Criterion scoring. In future research we would like to compare the performance of a larger data set of student compositions.

References

Attali, Y. (2004, April). Exploring the feedback and revision features of Criterion. Paper presented at the National Council on Measurement in Education, April 12, San Diego, CA.

Attali, Y., & Burstein, J. (November 2005). *Automated essay scoring with e-rater v.2.0*. ETS Research report.

Burstein, J. (2009). Opportunities for natural language processing research in education. Springer Lecture Notes in Computer Science, 5449, 6–27.

Burstein, J., Chodorow, M., & Leacock, C. (2003, August). Criterion: Online essay evaluation: An application for automated evaluation of student essays. Proceedings of the fifteenth annual conference on innovative applications of artificial intelligence. August 12, Acapulco, Mexico.

Burstein, J., & Wolska, M. (2003, April). Toward evaluation of writing style: Finding overly repetitive word use in student essays. In Proceedings of the 10th conference of the European chapter of the Association for Computational Linguistics, April 12, Budapest, Hungary.

Ericsson, P. F., & Haswell, R. (Ed.). (2006). *Machine scoring of student essays*. Logan, **UT**: Utah State University Press.
Hearst, M., Kukich, K., Hirschman, L., Breck, E., Light, M., Burge, J., Ferro, L., & Foltz, P. W. (2000). The debate on automated essay grading. ***IEEE*** *Intelligent Systems, 15(5).*
Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English. 63*(4), 480–499.
Jaschik, S. (2007). Fooling the College Board. *Inside Higher Education.* Retrieved May 22, 2011 from http://www.insidehighered.com/news/2007/03/26/writing
Jones, E. (2006). **ACCUPLACER**'s essay-scoring technology: When reliability does not equal validity. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays*. (93–113). Logan, **UT**: Utah State University Press.
McGee, T. (2006). Taking a spin on the Intelligent Essay Assessor. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays*. (79–92). Logan, **UT**: Utah State University Press.
Powers, D. E., Burstein, J., Chodorow, M., Fowles, M. E., & Kukich, K. (2000). Comparing the validity of automated and human essay scoring. ***GRE*** *Board Research.* 98(08a). Princeton, **NJ**: **ETS**.
Raimes, A. (1983). *Techniques in teaching writing*. New York, **NY**: Oxford University Press.
Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (in press). Automated essay scoring: Writing assessment and instruction. In E. Baker, B. McGaw & N. S. Petersen (Eds.), *International encyclopedia of education* (3rd ed.). Oxford, **UK**: Elsevier.
Shermis, M. D., & Burstein, J. (2002). *Automated essay scoring: A cross-disciplinary perspective.* Mahwah, **NJ**: Lawrence Erlbaum Associates.
Stone, S. (1995). *The primary multiage classroom: Changing schools for children.* Unpublished manuscript.
White, R. & Arndt, V. (1991). *Process writing*. London: Longman.
Winerip, M. (May 4, 2005). **SAT** essay test rewards length and ignores errors". New York Times. Retrieved May 22, 2011 from http://www.nytimes.com/2005/05/04/education/04education.html?pagewanted=print&position=

Appendix A

Criterion holistic scores

Score of 6:
You have put together a convincing argument. Here are some of the strengths evident in your writing:
Your essay:

- Looks at the topic from a number of angles and responds to all aspects of what you were asked to do
- Responds thoughtfully and insightfully to the issues in the topic
- Develops with a superior structure and apt reasons or examples (each one adding significantly to the reader's understanding of your view)
- Uses sentence styles and language that have impact and energy and keep the reader with you
- Demonstrates that you know the mechanics of correct sentence structure, and American English usage virtually free of errors

Score of 5:
You have solid writing skills and something interesting to say.
Your essay:

- Responds more effectively to some parts of the topic or task than to other parts
- Shows some depth and complexity in your thinking
- Organizes and develops your ideas with reasons and examples that are appropriate
- Uses the range of language and syntax available to you
- Uses grammar, mechanics, or sentence structure with hardly any error

Score of 4:
Your writing is good, but you need to know how to be more persuasive and more skillful at communicating your ideas.
Your essay:

- Slights some parts of the task
- Treats the topic simplistically or repetitively
- Is organized adequately, but you need more fully to support your position with discussion, reasons, or examples
- Shows that you can say what you mean, but you could use language more precisely or vigorously
- Demonstrates control in terms of grammar, usage, or sentence structure, but you may have some errors

Score of 3:
Your writing is a mix of strengths and weaknesses. Working to improve your writing will definitely earn you more satisfactory results because your writing shows promise.
In one or more of the following areas, your essay needs improvement. Your essay:

- Neglects or misinterprets important parts of the topic or task
- Lacks focus or is simplistic or confused in interpretation
- Is not organized or developed carefully from point to point
- Provides examples without explanation, or generalizations without completely supporting them
- Uses mostly simple sentences or language that does not serve your meaning
- Demonstrates errors in grammar, usage, or sentence structure

Score of 2:
You have work to do to improve your writing skills. You probably have not addressed the topic or communicated your ideas effectively. Your writing may be difficult to understand.
In one or more of the following areas, your essay:

- Misunderstands the topic or neglects important parts of the task
- Does not coherently focus or communicate your ideas
- Is organized very weakly or doesn't develop ideas enough
- Generalizes and does not provide examples or support to make your points clear
- Uses sentences and vocabulary without control, which sometimes confuses rather than clarifies your meaning

Score of 1:
You have much work to do in order to improve your writing skills. You are not writing

with complete understanding of the task, or you do not have much of a sense of what you need to do to write better. You need advice from a writing instructor and lots of practice. In one or more of the following areas, your essay:

- Misunderstands the topic or doesn't show that you comprehend the task fully
- Lacks focus, logic, or coherence
- Is undeveloped – there is no elaboration of your position
- Lacks support that is relevant
- Shows poor choices in language, mechanics, usage, or sentence structure which make your writing confusing

Score: N/A

There are areas of your writing that cannot be scored. Please check for errors in grammar, usage, punctuation and spelling.

Appendix B

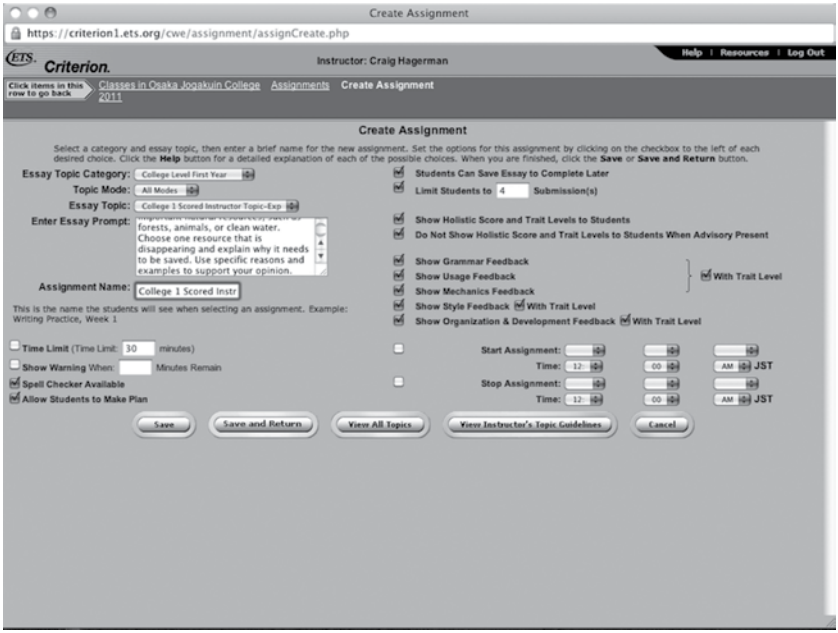
Criterion's trait feedback analysis categories

Category	Trait
Grammar	Fragment or missing comma
	Run-on sentences
	Garbled sentences
	Subject-verb agreement
	Ill-formed verbs
	Pronoun errors
	Possessive errors
	Wrong or missing word
	Proofread this!
Usage	Wrong article
	Missing or extra article
	Confused words
	Wrong form of word
	Faulty comparisons
	Preposition error
	Nonstandard word form
	Negation error
Mechanics	Spelling
	Capitalize proper nouns
	Missing initial capital letter in a sentence
	Missing question mark
	Missing final punctuation
	Missing apostrophe
	Missing comma
	Hyphen error
	Fused words

Category	Trait
Mechanics (ctd)	Compound words
	Duplicates
Style	Repetition of words
	Inappropriate words or phrases
	Sentences beginning with coordinating conjunctions
	Too many short sentences
	Too many long sentences
	Passive voice
	Number of words
	Number of sentences
	Average number of words per sentence
Organization & Development	Thesis Statement
	– Topic relationship & technical quality
	Main ideas
	Supporting ideas
	Conclusion
	Transitional words and phrases

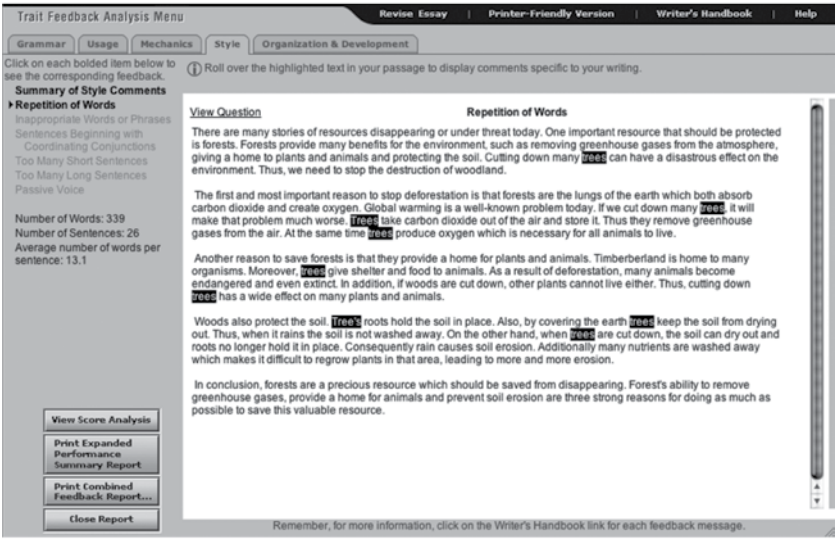
Appendix C

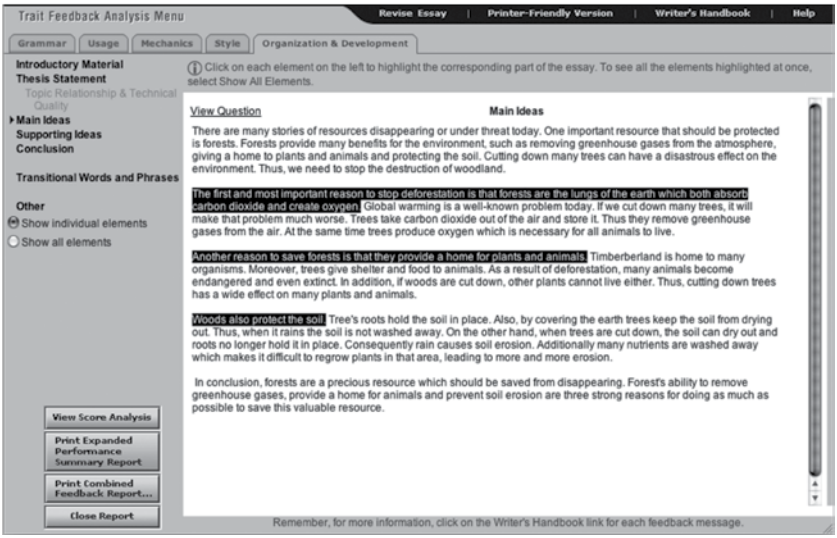
Criterion assignment options

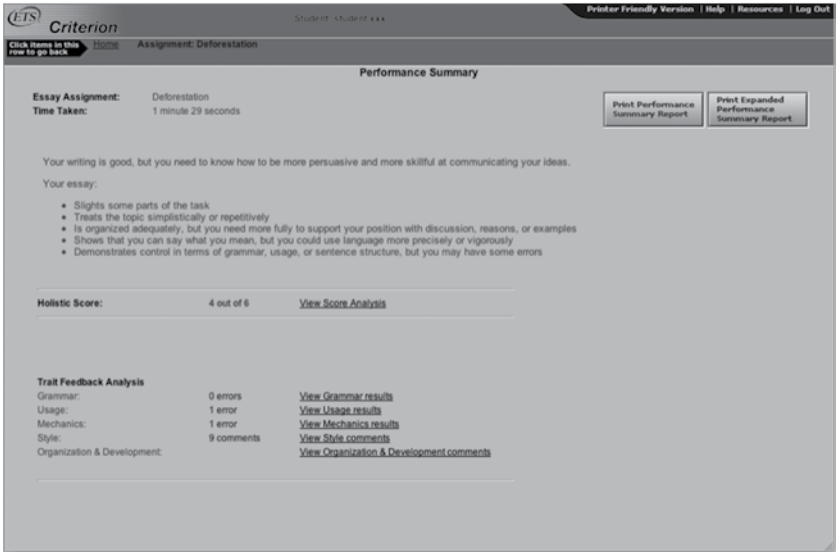


Appendix D

Criterion trait analysis feedback






ETS Criterion
Student student xxx
Printer Friendly Version | Help | Resources | Log Out
Click items in this row to go back
Home
Assignment: Deforestation
Performance Summary
Essay Assignment: Deforestation
Time Taken: 1 minute 29 seconds
Print Performance Summary Report
Print Expanded Performance Summary Report
Your writing is good, but you need to know how to be more persuasive and more skillful at communicating your ideas.
Your essay:
Slights some parts of the task
Treats the topic simplistically or repetitively
Is organized adequately, but you need more fully to support your position with discussion, reasons, or examples
Shows that you can say what you mean, but you could use language more precisely or vigorously
Demonstrates control in terms of grammar, usage, or sentence structure, but you may have some errors
Holistic Score: 4 out of 6 View Score Analysis
Trait Feedback Analysis
Grammar: 0 errors View Grammar results
Usage: 1 error View Usage results
Mechanics: 1 error View Mechanics results
Style: 9 comments View Style comments
Organization & Development: View Organization & Development comments