# Exploring the Feedback and Revision Features of *Criterion*

Yigal Attali

ETS, Princeton, NJ

Paper presented at the National Council on Measurement in Education (NCME)
held between
April 12 to 16, 2004, in San Diego, CA.


## Abstract

*Criterion*[SM] on-line essay evaluation application provides students with immediate feedback about their essay with Critique writing analysis tools. The feedback is related to grammar, usage, mechanics, style, and organization and development. The major research question in this study is whether the feedback report is helpful for students in subsequent revisions of their essays. Helpful feedback can guide students in revising an essay, and will result in a better feedback report for the essay in subsequent re-submissions (i.e., one that has fewer critical comments). A positive feedback report has significant instructional implications and it also supports the validity of *Criterion,* since an iterative feedback and revision process is natural to the writing process. *Criterion* provides a way of automating and speeding up this process with its diagnostic feedback and essay scoring capabilities.

## Introduction

Revision is a central process in cognitive models of writing (Hayes & Flower, 1980; Hayes, 1996; Bereiter & Scardamalia, 1987; Butterfield, Hacker, & Albertson, 1996) that emphasizes the writing process, in addition to the writing product. Based on several theoretical treatments of the writing process, Fitzgerald (1987) gave the following definition of revision (p. 484):

> Revision means making any changes at any point in the writing process. It involves identifying discrepancies between intended and instantiated text, deciding what could or should be changed in the text and how to make desired changes, and operating, that is, making the desired changes. Changes may or may not affect meaning of the text, and they may be major or minor. Also, changes may be made in the writer's mind before being instantiated in written text, at the time text is first written, and/or after text is first written.

Research on revision found that, especially for high-school age and older or more skilled writers, revisions appear to improve the quality of compositions (see, Fitzgerald, 1987 for a review). The importance of feedback was also studied in the context of revision in writing. There is evidence that teacher or peer feedback can enhance revision for writers in the primary grades through high school, especially if the feedback is focused and part of a wider instructional program (Gere & Stevens, 1985; Kamler, 1980). Younger students, and even many older students, do very little revision without peer group or teacher support (Scardamalia & Bereiter, 1986).

The effects of computers on writing have been an increasingly important subject of research in the past 30 years. Even the simplest word-processor allows users to make changes to text that would facilitate the revision process, and would have been more cumbersome on paper. Indeed, three meta-analyses about the effect of computers on student writing found that writing, using computers, has a significant positive effect on writing quantity, quality, and revision (Bangert-Drowns, 1993; Cochran-Smith, 1991; Goldberg, Russell, & Cook, 2003).

However, the potential of the computer as a cognitive tool is greater than allowing "first-order fingertip effects" (Perkins, 1985), that is putting more power at the user's fingertips. Particularly, natural language processing (NLP) technologies have the power to analyze and provide automated feedback to writers in ways that might further enhance the interactive process

of writing and the promotion of better understanding of best practices for writing. Work in automated feedback was initiated with the Writer's Workbench (MacDonald et al., 1982) and several reports on other systems were published (e.g., Holdich & Chung, 2003; Zellermayer, Salomon, Globerson, & Givon, 1991).

### *Criterion*

*Criterion*[SM] is a web-based service developed by ETS to evaluate a student's writing skill and provide instantaneous score reporting and diagnostic feedback. (For a detailed description of the system, see Burstein, Chodorow, and Leacock (2003).) *Criterion* contains two complementary applications that are based on NLP methods. The scoring application, e-rater, extracts linguistically-based features from an essay and uses a statistical model to determine how these features are related to overall writing quality, so that a holistic score may be assigned to the essay. The second application, *Critique,* is comprised of a suite of programs that evaluates and provides feedback for errors in grammar, usage, and mechanics, identifies the essay's discourse structure, and recognizes undesirable stylistic features.

The writing analysis tools identify five main types of grammar, usage, and mechanics errors – agreement errors, verb formation errors, wrong word use, missing punctuation, and typographical errors. The approach to detecting violations of general English grammar is corpus based and statistical, and can be explained as follows. The system is trained on a large corpus of edited text, from which it extracts and counts sequences of adjacent word and part-of-speech pairs called bigrams. The system then searches student essays for bigrams that occur much less often than would be expected based on the corpus frequencies (Chodorow & Leacock, 2000). The writing analysis tools also highlight aspects of style that the writer may wish 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, a property of the essay that might affect its rating of overall quality (Burstein & Wolska, 2003). Finally, the writing analysis tools provide feedback about discourse elements present or absent in the essay. A well-written essay should contain discourse elements, which include introductory material, a thesis statement, main ideas, supporting ideas, and a conclusion. In order to identify the various discourse elements, the system was trained on a large corpus of human annotated essays (see Burstein, Marcu, and Knight, 2003).

### *Purpose of the Study*

During the 2002-2003 school year *Criterion* was used by thousands of students nationally in 6th through 12th grade. The purpose of this study was to evaluate the effectiveness of the automated feedback and revision features of *Criterion* by focusing on the (possible) improvement in feedback from first to last submission of an essay.[1] This is the simplest question one could ask about a system that provides automated feedback to the student: Are students capable of using the feedback given to them to correct the errors detected? Negative findings could be explained by difficulties in understanding the feedback, difficulties in understanding how to correct the errors, or even motivation problems. However, positive results are a necessary to support claims about instructional usefulness of the system over time.

This study involves an analysis of data collected from a large-scale production implementation of *Criterion*. Consequently, the degree of control and knowledge about students' responses was limited. For instance, there was no background information available about students, and as was mentioned above, intermediate submissions could not be analyzed. The only information available was the essay text of the first and last submission together with the scores and feedback report for these essays, the number of submissions (the submission number of the last submission), and the grade for which the essay prompt was designed.

## Results

### *Description of Dependent Variables*

In addition to the e-rater score as an overall measure of essay quality, essay length (measured as number of words) was also included because it is generally highly correlated with both human and automated essay scores. *Criterion* identifies and computes 33 different grammar, usage, mechanics, and style errors. The counts of these errors were transformed into rates by dividing them by the essay length. In addition to the individual error rates, an overall measure of grammar, usage, mechanics, and style errors was computed by summing the individual error rates.

[1] Due to issues of data storage intermediate submissions were not saved in the database.

The organization and development module of *Criterion* identifies background (introductory material), thesis, main points, supporting ideas, and conclusion discourse elements in the essays (in addition, it also labels elements as *Other* if it was not able to recognize them). Analyses for this study used a count of the different elements (an element is the longest consecutive number of sentences assigned to one discourse category) with the exception of supporting ideas elements that were counted only when they immediately followed a main point element, and main point elements that were restricted to three different elements per essay.

These restrictions follow the five-paragraph essay strategy for developing writers that was adopted in *Criterion*. According to this strategy novice writers should typically include in their essay an introductory paragraph, a three-paragraph body (three paragraphs, each containing a single main point/supporting idea pair), and a concluding paragraph.

This approach makes it possible to compute, in addition to the counts of individual discourse elements, an overall development score by summing up the counts of the thesis, main points, supporting ideas, and conclusion elements as defined above. In this study, the development score was defined as the sum of the development elements – 8. This development score may be interpreted as the difference between the actual and optimal development. A score of –8 means that there are no required elements, whereas a score of 0 means that all required elements (thesis, conclusion, three main points and corresponding supporting ideas) are present and there is no discrepancy between optimal and existent development.

***General Description of the Dataset***

The dataset included 33,171 student essay submissions with more than 50 words written to *Criterion* prompts. This minimum essay length was chosen to exclude submissions of clearly non-motivated students. Of these, 23,567 (71%) were submitted only once and 9,604 were submitted more than once. This in itself suggests that most students did not exploit the revision capabilities of the *Criterion* system. For the essays that were submitted more than once, the first and last submission were available for analysis. For this study, it was decided to include only essays that were submitted 10 times or less because the number of essays for higher number of submissions was low (<100). In this way 9,275 essays were retained out of the original number of 9,604 (97%). There were 4461, 2056, 1036, 612, 423, 290, 179, 115, and 103 essays for 2 to 10 submissions. Recall that the student report does retain the total number of submissions,

though it does not store the text of all submissions. The essays were written to 38 different prompts (there were between 157 and 2,662 essays per prompt), designed for 6th to 12th grade. The number of essays per grade was 1012, 1014, 1624, 1451, 2162, 1337, and 675 for grades 6 to 12, respectively.

### *Overall Differences Between Revised and Unrevised Essays*

Table 1 shows the differences in e-rater scores and essay length (in number of words) for the essays that were not revised and for those that were. E-rater scores are scaled to a whole number from 1 to 6, in the same way human holistic scores are scaled. The Table shows that the first submissions of revised essays were shorter and received lower scores than unrevised essays, but the differences were rather small: both effect sizes *d* are equal to .18, whereas Cohen (1988) defines a "small" effect size as .2 or larger.

**Table 1**

***Descriptive Statistics for Revised and Unrevised Essays***

| | Unrevised essays (*N* = 23,567) | | First submission of revised essays (*N* = 9,275) | |
|---|---|---|---|---|
| | Mean | STD | Mean | STD |
| e-rater score | 3.90 | 1.15 | 3.70 | 1.09 |
| Essay length (words) | 286 | 146 | 260 | 143 |

### *Changes in Scores and Essay Length From First to Last Submission*

The essay revisions had significant effects on major dependent variables. Table 2 shows that e-rater scores improved by almost half the standard deviation of the scores in the first submission (an effect size of .47) and that essay length increased by an average of .39 of the standard deviation of first submission essay lengths. Development scores improved by .31 of the standard deviation and error rates decreased by .15 to .27 of the respective standard deviations for grammar, usage, mechanics, and style errors.

**Table 2**

***Descriptive Statistics for Major Measures***

| | Mean in first submission | STD in first submission | Difference between last and first sub. | STD of difference | Effect size |
|---|---|---|---|---|---|
| e-rater score | 3.70 | 1.09 | 0.51* | 0.89 | 0.47 |
| Essay length (words) | 260 | 143 | 55* | 106 | 0.39 |
| Development | -3.07 | 2.56 | 0.79* | 1.91 | 0.31 |
| Grammar | 0.0005 | 0.0008 | -0.0001* | 0.0007 | -0.15 |
| Usage | 0.0005 | 0.0009 | -0.0001* | 0.0007 | -0.16 |
| Mechanics | 0.0020 | 0.0027 | -0.0006* | 0.0025 | -0.21 |
| Style | 0.0186 | 0.0142 | -0.0038* | 0.0110 | -0.27 |

*Note.* Effect size is defined as difference divided by the standard deviation of first submission.
*The Wilcoxon signed-ranks test was significant at the .01 level, two-tailed.

***Improvements in Individual Feedback Measures***

Table 3 shows the extent of the different errors in the first and last submission. The first column shows the percent of essays in the first submission with at least one error of each type. Subsequent columns relate only to essays that had errors in the first or last submission. The second and third column shows the mean and standard deviation of the error rate in the first submission. The fourth column shows the mean difference in error rate between the last and first submission. A negative difference is expected if the feedback has a positive impact. The differences that were *not* significant at the .01 significance level (for a Wilcoxon signed-ranks test) are marked with an asterisk. The fifth column presents the effect size of the difference in error rates, defined as the mean difference divided by the standard deviation of the error rates for the first submission (the standard deviations for the first and last submissions are very similar). The last column presents another effect size measure, the mean of the percent decrease of the error rates, defined as the difference in error rate divided by the error rate for the first submission. Ideally, the percent decrease would be 100%.

The main conclusions from the Table are as follows. The extent of the different types of errors varied considerably. Three of the error types were not found at all in the analyzed dataset,

and, on the other hand, spelling errors and repetition of words were found in 78% and 93% of the essays.

For twenty-three out of thirty error types that were found in the essays, there was a significant decrease in the error rates from first to final submissions. Of the non-significant differences, six were associated with rare errors (up to 3% of essays included these errors) and only one common error (found in 16% of essays), the advice for "too many short sentences", was non-significant.

Sixteen of the thirty error types showed "small" effect sizes, one showed a "medium" (defined by Cohen, 1988, as .5-.8) effect size (garbled sentences), and thirteen error types showed smaller effect sizes. The median effect size was .22 and the median percent decrease in errors was 24%, or, in other words, about a quarter of the errors were corrected in the final version.

In conclusion, it seems that students were sensitive to the feedback for most error types and were able to correct errors in subsequent versions of their essays.

**Table 3**

***Descriptive Statistics of Grammar, Usage, Mechanics, and Style Error Rates***

| | Percent of essays with errors | Mean error rate in first submission | STD of error rate in first submission | Difference in error rates | Effect size | Mean percent decrease |
|---|---|---|---|---|---|---|
| **Grammar** | | | | | | |
| Fragments | 35% | 0.0051 | 0.0061 | -0.0010 | -0.17 | 20% |
| Run-On Sentences | 0% | . | . | . | . | . |
| Garbled Sentences | 2% | 0.0042 | 0.0036 | -0.0020 | -0.55 | 48% |
| Subject-Verb Agreement | 19% | 0.0040 | 0.0046 | -0.0007 | -0.16 | 19% |
| Ill-formed Verbs | 11% | 0.0035 | 0.0036 | -0.0009 | -0.25 | 25% |
| Pronoun Error | <1% | 0.0029 | 0.0038 | 0.0005* | 0.14 | -18% |
| Missing Possessive Error | 21% | 0.0039 | 0.0040 | -0.0010 | -0.26 | 27% |
| Wrong or Missing Word | 3% | 0.0032 | 0.0035 | -0.0004* | -0.12 | 14% |
| Proofread This! | 15% | 0.0036 | 0.0042 | -0.0009 | -0.22 | 25% |
| **Usage** | | | | | | |
| Wrong Article | 7% | 0.0034 | 0.0035 | -0.0008 | -0.22 | 22% |
| Missing Article | 0% | . | . | . | . | . |
| Confused Words | 48% | 0.0066 | 0.0069 | -0.0019 | -0.27 | 28% |
| Wrong Form of Word | 1% | 0.0032 | 0.0033 | -0.0009* | -0.27 | 28% |
| Faulty Comparisons | 1% | 0.0034 | 0.0029 | -0.0002* | -0.07 | 6% |
| Preposition Error | 0% | . | . | . | . | . |
| Nonstandard Verb or Word Form | 1% | 0.0043 | 0.0051 | -0.0012 | -0.23 | 27% |
| **Mechanics** | | | | | | |
| Spelling | 78% | 0.0192 | 0.0236 | -0.0051 | -0.22 | 27% |
| Capitalize Proper Nouns | 21% | 0.0091 | 0.0116 | -0.0034 | -0.29 | 38% |
| Missing Initial Capital Letter in a Sentence | 21% | 0.0088 | 0.0125 | -0.0019 | -0.15 | 22% |
| Missing | 4% | 0.0035 | 0.0040 | -0.0006 | -0.15 | 17% |

| | | | | | | |
|---|---|---|---|---|---|---|
| Question Mark | | | | | | |
| Missing Final Punctuation | 11% | 0.0047 | 0.0065 | -0.0015 | -0.23 | 32% |
| Missing Apostrophe | 15% | 0.0069 | 0.0077 | -0.0024 | -0.31 | 35% |
| Missing Comma | 13% | 0.0035 | 0.0033 | -0.0007 | -0.21 | 20% |
| Hyphen Error | 7% | 0.0039 | 0.0039 | -0.0010 | -0.25 | 25% |
| Fused Words | 8% | 0.0054 | 0.0057 | -0.0020 | -0.34 | 36% |
| Compound Words | 12% | 0.0039 | 0.0038 | -0.0009 | -0.24 | 23% |
| Duplicates | 9% | 0.0037 | 0.0069 | -0.0012 | -0.17 | 32% |
| **Style** | | | | | | |
| Repetition of Words | 93% | 0.1146 | 0.0803 | -0.0248 | -0.31 | 22% |
| Inappropriate Words or Phrases | 1% | 0.0068 | 0.0406 | -0.0029* | -0.07 | 43% |
| Too Many Sentences Beginning with Coord. Conj. | 2% | 0.0113 | 0.0112 | 0.0016* | 0.15 | -14% |
| Too many short sentences | 16% | 0.0225 | 0.0234 | 0.0002* | 0.01 | -1% |
| Too many long sentences | 5% | 0.0037 | 0.0033 | -0.0004 | -0.12 | 11% |
| Passive Voice | 13% | 0.0033 | 0.0030 | -0.0003 | -0.11 | 10% |

*Note.* Effect size is defined as difference in error rates divided by the standard deviation of error rates in first submission. Mean percent decrease is defined as the difference in error rates divided by the error rate for first submission.
*The Wilcoxon signed-ranks test was *not* significant at the .01 level, two-tailed.

Table 4 presents the results for the identification of discourse elements in the student essays. We should expect that the occurrence of the background, thesis, and conclusion elements would be higher in the last submission, that the number of main points and supporting ideas would increase too, whereas the number of *Other* elements would decrease. The Table shows that these expectations were met for the background, main point, supporting ideas, and conclusion elements, but not for the thesis and *Other* elements. For the *Other* element a very small but significant *increase* in occurrence was found. This element includes titles and opening and closing salutations, which might explain this result. For the thesis statement the difference

in occurrence between the first and last submissions was not significant. For the other four types of elements the increase in occurrence was significant with small effect sizes (.18 to .34).

**Table 4**

***Descriptive Statistics for the Discourse Elements***

| Element | Range of values | Mean in first submission | STD in first submission | Difference between last and first sub. | Effect size |
|---|---|---|---|---|---|
| Background | 0-1 | 0.55 | 0.50 | 0.09 | 0.18 |
| Thesis | 0-1 | 0.79 | 0.40 | -0.01* | -0.03 |
| Main-Point | 0-3 | 1.78 | 1.13 | 0.34 | 0.30 |
| Supporting Ideas | 0-3 | 1.76 | 1.13 | 0.34 | 0.30 |
| Conclusion | 0-1 | 0.60 | 0.49 | 0.14 | 0.28 |
| Other | 0- | 0.32 | 0.47 | 0.02 | 0.04 |

*Note.* Effect size is defined as difference divided by the standard deviation of first submission.
*The Wilcoxon signed-ranks test was *not* significant at the .01 level, two-tailed.

***Analyses of Covariance***

The purpose of the next set of analyses was to assess the effects of feedback in the context of two independent variables, number of submissions and grade level, while taking into account initial and final essay length. It was expected that the effect of feedback on grammar, usage, mechanics, style, and development scores would be greater for students who submitted more versions of their essays. On the other hand, there was no reason to expect that the effects would differ for different grade levels ($6^{th}$ to $12^{th}$ in this study). To answer these questions a 9 X 7 X 2 mixed between and repeated measures analysis of covariance was performed for each of the five measures. The between-subjects independent variables were number of submissions (from 2 to 10) and grade level ($6^{th}$ to $12^{th}$). The repeated measures independent variable was submission (first and last). The covariates were initial and final essay length.

Tables 5-9 present the main results of the ANOVA. The first question addressed in this analysis is whether there is an effect of feedback on scores. Do the first and last submission, represented here as the within-subjects independent variable of Feedback, elicit the same average scores (development or the various error rates), independent of groups? In the jargon of profile analysis this is the test of the "flatness" hypothesis. The Tables show that in all cases the

feedback effect is highly significant. The effects shown take into account the effects of the two covariates.

Although the first question is the main interest of this analysis, it cannot be answered independently of a second important question, that of "parallelism" of profiles. Do different groups have parallel profiles for the two dependent variables? In this context the test of parallelism is the test of interaction between the within-subject independent variable and the between-subject independent variable. The reason that the first question of flatness is dependent on the second question of parallelism is that if there is an interaction, i.e., the profiles are not parallel, so there must be at least one group with a profile that is not flat. There are two parallelism tests in this study, one for each between-subject independent variable, and there are different predictions for these tests.

In the case of the number of submissions a significant interaction is expected because we do not expect the first submission to differ in scores as the number of submissions go up, but we do expect an increasing effect on the last submission as the number of submissions goes up. If this expectation were correct an interaction would be found. In fact, these specific hypotheses can be interpreted as two other contrast hypotheses. We expect no trend effect of submission on the scores of the first submission but we do expect a trend effect on the last submission. The simplest kind of trend is a linear trend. The non-parallel hypothesis and the two contrast hypotheses are generally supported with all five kinds of scores. We find significant interactions between feedback and submissions, non-significant linear effect of submission for the first score, and a significant linear effect of submission for the last score, except for the grammar score where the effect is almost not significant.

The linear trend of the different final scores is presented in Figure 1, where the final submission least-square mean scores (adjusted for both covariates), for each number of submissions, are presented. In order to facilitate a comparison between different scores that have different typical values, these means were "standardized" by dividing each mean by the mean the score for the group with a total of two submissions. The Figure shows the general improvement in scores as the number of submissions increases. The improvement is perfectly monotonic for two to five submissions, where 88% of the observations are concentrated (the apparent "decrease" in development scores is due to its "negative" definition as number of lacking elements, so a decrease in this number constitutes an improvement).

In the case of the second between-subject independent variable, grade level, the expectation is of parallel profiles, or no interaction between feedback and grade level. In other words, different feedback effects are not expected for different grades. For two of the scores, usage and mechanics, a small significant interaction was found, but post-hoc analyses revealed no monotonic trend.

The third kind of question that can be answered in profile analysis is the "levels" hypothesis. Do different groups score differently, on average, on the collected set of dependent measures (first and last submission scores)? This is the question of the between-subject main effect in regular ANOVA. In the case of the number of submissions independent variable, this question is not important, since for the first submission we are not expecting systematic differences whereas for the last submission we do. On the other hand, we do expect an effect of grade in the overall level of scores, more specifically, a decrease in the rate of errors and an increase in the level of development with grade. This hypothesis was supported in all cases.

Lastly, the interaction between the within-subject feedback score and the two covariates evaluates the usefulness of the covariates in adjusting the means of the scores. In three cases, the development, mechanics, and style scores, the interactions were significant and indicate that the covariates are important predictors of feedback improvement in the context of the other independent variables. For the other two scores, grammar and usage, the covariates were not found to adjust the dependent scores in a significant way.

**Table 5**

***ANOVA Summary for Development Score***

| Type of effect | Source of Variance | df | SS | MS | *F* | Pr > *F* |
|---|---|---|---|---|---|---|
| Flatness | Feedback | 1 | 433.5 | 433.5 | 348.9 | <.0001 |
| Parallelism | Feedback*Submission | 8 | 46.8 | 5.8 | 4.7 | <.0001 |
| Parallelism | Feedback*Grade | 6 | 10.9 | 1.8 | 1.5 | 0.1865 |
| Levels | Submission | 8 | 129.3 | 16.2 | 2.3 | 0.0199 |
| Levels | Grade | 6 | 437.0 | 72.8 | 10.3 | <.0001 |
| CV Utility | Feedback*FEL | 1 | 4806.4 | 4806.4 | 3868.3 | <.0001 |
| CV Utility | Feedback*LEL | 1 | 3694.9 | 3694.9 | 2973.8 | <.0001 |
| Linear contrasts | | | | | | |
| First score | Submission | 1 | 11.5 | 11.5 | 2.7 | 0.1021 |
| Last score | Submission | 1 | 83.4 | 83.4 | 20.6 | <.0001 |

*Note.* FEL is first essay length, LEL is last essay length.

**Table 6**

***ANOVA Summary for Grammar Score***

| Type of effect | Source of Variance | df | SS | MS | *F* | Pr > *F* |
|---|---|---|---|---|---|---|
| Flatness | Feedback | 1 | 0.00001061 | 0.00001061 | 44.2 | <.0001 |
| Parallelism | Feedback*Submission | 8 | 0.00000695 | 0.00000087 | 3.6 | 0.0003 |
| Parallelism | Feedback*Grade | 6 | 0.00000228 | 0.00000038 | 1.6 | 0.1467 |
| Levels | Submission | 8 | 0.00001246 | 0.00000156 | 2.0 | 0.0383 |
| Levels | Grade | 6 | 0.00015453 | 0.00002576 | 33.7 | <.0001 |
| CV Utility | Feedback*FEL | 1 | 0.00000070 | 0.00000070 | 2.9 | 0.0881 |
| CV Utility | Feedback*LEL | 1 | 0.00000033 | 0.00000033 | 1.4 | 0.2404 |
| Linear contrasts | | | | | | |
| First score | Submission | 1 | 0.00000008 | 0.00000008 | 0.1 | 0.7039 |
| Last score | Submission | 1 | 0.00000159 | 0.00000159 | 3.6 | 0.0583 |

*Note.* FEL is first essay length, LEL is last essay length.

**Table 7**

***ANOVA Summary for Usage Score***

| Type of effect | Source of Variance | df | SS | MS | *F* | Pr > *F* |
|---|---|---|---|---|---|---|
| Flatness | Feedback | 1 | 0.00002472 | 0.00002472 | 95.9 | <.0001 |
| Parallelism | Feedback*Submission | 8 | 0.00001652 | 0.00000206 | 8.0 | <.0001 |
| Parallelism | Feedback*Grade | 6 | 0.00000406 | 0.00000068 | 2.6 | 0.0152 |
| Levels | Submission | 8 | 0.00002115 | 0.00000264 | 2.7 | 0.0057 |
| Levels | Grade | 6 | 0.00002437 | 0.00000406 | 4.2 | 0.0004 |
| CV Utility | Feedback*FEL | 1 | 0.00000052 | 0.00000052 | 2.0 | 0.1545 |
| CV Utility | Feedback*LEL | 1 | 0.00000055 | 0.00000055 | 2.2 | 0.1426 |
| Linear contrasts | | | | | | |
| First score | Submission | 1 | 0.00000015 | 0.00000015 | 0.2 | 0.6519 |
| Last score | Submission | 1 | 0.00000843 | 0.00000843 | 17.0 | <.0001 |

*Note.* FEL is first essay length, LEL is last essay length.

**Table 8**

***ANOVA Summary for Mechanics Score***

| Type of effect | Source of Variance | df | SS | MS | *F* | Pr > *F* |
|---|---|---|---|---|---|---|
| Flatness | Feedback | 1 | 0.00033119 | 0.00033119 | 103.8 | <.0001 |
| Parallelism | Feedback*Submission | 8 | 0.00011415 | 0.00001427 | 4.5 | <.0001 |
| Parallelism | Feedback*Grade | 6 | 0.00004293 | 0.00000715 | 2.2 | 0.0364 |
| Levels | Submission | 8 | 0.00050276 | 0.00006284 | 6.4 | <.0001 |
| Levels | Grade | 6 | 0.00032513 | 0.00005419 | 5.5 | <.0001 |
| CV Utility | Feedback*FEL | 1 | 0.00022619 | 0.00022619 | 70.9 | <.0001 |
| CV Utility | Feedback*LEL | 1 | 0.00013850 | 0.00013850 | 43.4 | <.0001 |
| Linear contrasts | | | | | | |
| First score | Submission | 1 | 0.00000106 | 0.00000106 | 0.2 | 0.6939 |
| Last score | Submission | 1 | 0.00006973 | 0.00006973 | 11.3 | 0.0008 |

*Note.* FEL is first essay length, LEL is last essay length.

**Table 9**

***ANOVA Summary for Style Score***

| Type of effect | Source of Variance | df | SS | MS | *F* | Pr > *F* |
|---|---|---|---|---|---|---|
| Flatness | Feedback | 1 | 0.02029631 | 0.02029631 | 388.6 | <.0001 |
| Parallelism | Feedback*Submission | 8 | 0.00482805 | 0.00060351 | 11.6 | <.0001 |
| Parallelism | Feedback*Grade | 6 | 0.00042825 | 0.00007138 | 1.4 | 0.224 |
| Levels | Submission | 8 | 0.00610007 | 0.00076251 | 3.6 | 0.0004 |
| Levels | Grade | 6 | 0.15345203 | 0.02557534 | 119.7 | <.0001 |
| CV Utility | Feedback*FEL | 1 | 0.06034116 | 0.06034116 | 1155.3 | <.0001 |
| CV Utility | Feedback*LEL | 1 | 0.03863721 | 0.03863721 | 739.7 | <.0001 |
| Linear contrasts | | | | | | |
| First score | Submission | 1 | 0.00001782 | 0.00001782 | 0.1 | 0.7248 |
| Last score | Submission | 1 | 0.00450666 | 0.00450666 | 36.9 | <.0001 |

*Note.* FEL is first essay length, LEL is last essay length.

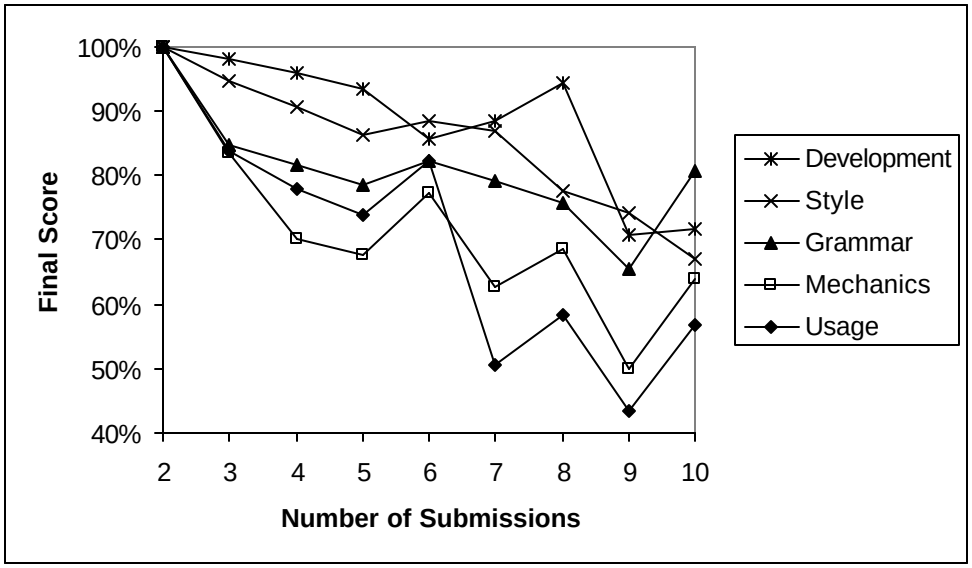


**Figure 1.** ***Least-Square (Adjusted for Both Covariates) Mean Feedback Scores for Different Number of Submissions, Expressed as Percent of Mean Score for Two Submissions***

***Relation Between Feedback Improvement Scores and Initial Essay Quality***

Examining the differential effect of feedback on low and high quality essays is important for the evaluation of *Criterion*. It is important to know whether low or high quality essays benefit more from the feedback provided to the students. Table 10 presents the correlations between the improvement scores and two measures of (first) essay quality, e-rater score and essay length. As expected from the feedback-covariate interaction results above, only the development, mechanics, and style scores had significant correlations. In all of these cases a negative relation was found, or, in other words, there is a tendency for higher improvement in development, mechanics, and style, with lower initial scores.

**Table 10**

***Correlations Between Major Improvement Scores and Between First Essay Score and Essay Length***

| Difference score | First e-rater score | First essay length |
|---|---|---|
| Development | -.30* | -.29* |
| Grammar | .02 | .01 |
| Usage | -.01 | .00 |
| Mechanics | -.06* | -.06* |
| Style | -.23* | -.21* |

* $p < .01$

Figure 2 presents, visually, the relation between first essay length and improvement in feedback scores. Essay length values were rounded to the nearest whole multiple of 100 (the Figure shows the results for up to a rounded essay length of 600 words, covering 99% of the essays). The improvement scores were transformed in a similar way to Figure 1, where the mean improvement of each group was divided by the mean of the first group, in this case the improvement scores for essays with a rounded length of 100. The Figure shows that mechanics, style, and development improvement scores are monotonically decreasing, whereas grammar and usage improvement scores increase, and then decrease with essay length.

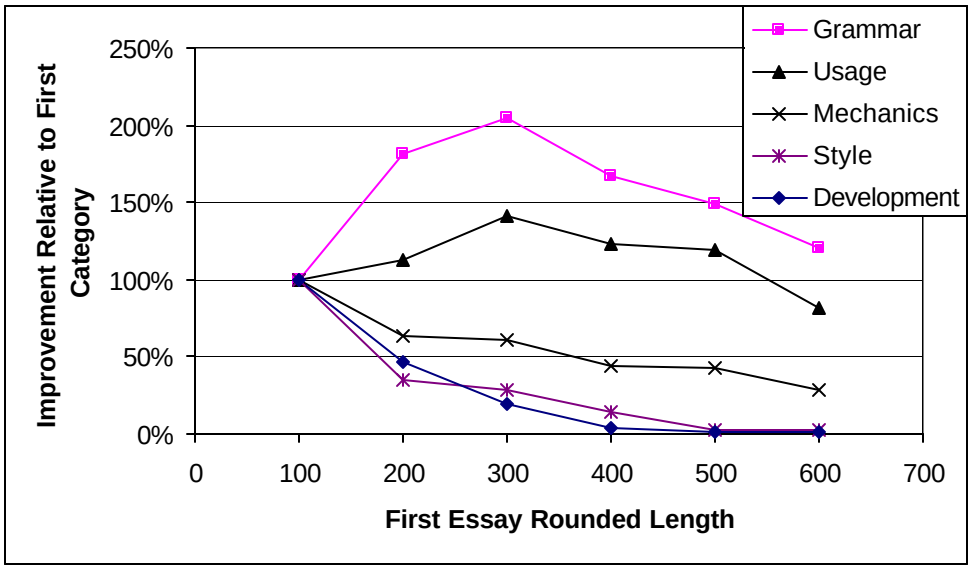


**Figure 2.** ***Mean Improvement Scores Expressed as Percent of Mean Score for a Rounded Essay Length of 100***

## Summary and Further Research

The *Criterion* system provides automatic writing feedback following the submission of an essay by students. The feedback covers five major areas of writing quality: organization and development, grammar, usage, mechanics, and style. This study investigated the improvement in writing feedback that students receive after resubmitting an essay more than once. The overall purpose was to assess whether students understand the feedback provided to them and have the ability to attend to the comments the system provides to them.

This investigation was conducted in the context of a large-scale field implementation of the system. More than 9,000 essays that were submitted more than once by six to twelve grade students were identified and used in the study. Both major and minor feedback aspects were compared from first to last submission, taking into account the number of submissions, the grade level, and overall quality of the essay measured by the e-rater score and essay length.

Results showed that students were able to significantly lower the rate of most of the 30 specific error types that were identified by the system and reduced their error rates by about one quarter (with a median effect size of .22). With respect to feedback about specific discourse elements in the essay, students were able to significantly increase the rate of occurrence of background and conclusion elements (but not thesis statement) and significantly increased the

number of main points and supporting ideas elements, following feedback. Students were not able to decrease the rate of elements not recognized by the system (other elements).

In terms of overall feedback scores, students improved their development scores by .31 of a standard deviation, and improved relatively less their grammar, usage, mechanics, and style scores (.15, .16, .21, and .27 of a standard deviation, respectively). The length of the student essays increased relatively more than the feedback scores, .39 of a standard deviation; and the e-rater score increased by almost half of a standard deviation from first to last submission. These results are an indication of the effectiveness of changes made by students following feedback.

An analysis of covariance on the major improvement scores from first to last submission showed that, after controlling for first and last essay lengths, the five improvement scores were significant, there was a general linear increase in the improvement with increasing submissions, and there was no coherent pattern for different grades.

An analysis of the relation between improvement and initial essay quality (first e-rater score and first essay length) showed that in the case of mechanics, style, and development scores, improvement tended to decrease for longer essays (that are generally better essays), whereas in the case of grammar and usage, improvement tended to be highest for middle length essays.

In sum, the results show that students are able, to some significant extent, to understand and attend to the feedback provided in *Criterion*. Of particular importance is their ability to improve the development of the essay beyond the mere lengthening of it. The results of this study could help to improve the system by examining areas where students were not able to improve the feedback. For this purpose a micro-level analysis of particular error types might be needed. For example, what kinds of missing or wrong articles are corrected and which are not? Another interesting analysis would investigate the ways students attended to development feedback. What were the strategies students used to more fully develop their essays?

The instructional utility of *Criterion* feedback can also be studied to some extent in this context of deployed implementation, by analyzing submissions for different prompts by the same students and comparing specific error rates of students who corrected these errors in previous essay submissions.

## References


Bangert-Drowns, R. L. (1993). The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. *Review of Educational Research, 63,* 69–93.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Erlbaum.

Burstein, J., Chodorow, M., and Leacock, C. (2003). *Criterion*[SM]: Online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*.

Burstein, J., & Wolska, M. (2003). Toward evaluation of writing style: Overly repetitious word use in student writing. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*. Budapest, Hungary.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems: Special Issue on Natural Language Processing* 18(1): 32-39.

Butterfield, E.C., Hacker, D.J., & Albertson, L.R. (1996). Environmental, cognitive and metacognitive influences on text revision: Assessing the evidence. *Educational Psychology Review, 8,* 239-297.

Chodorow, M., & Leacock, C. (2000). An unsupervised method for detecting grammatical errors. *Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, 140-147.

Cochran-Smith, M. (1991). Word processing and writing in elementary classrooms: A critical review of related literature. *Review of Educational Research, 61,* 107–155.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research, 57*, 481-506.

Gere, A.R., & Stevens, R.S. (1985). The language of writing groups: How oral response shapes revision. In S. W. Freedman (Ed.), *The acquisition of written language: Response and revision* (pp. 85-105). Norwood, NJ: Ablex.

Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A meta-analysis of studies from 1992 to 2002. *Journal of Technology, Learning, and Assessment, 2*(1).

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Randsdell (Eds.), *The science of writing* (pp. 1-27). Mahwah, NJ: Erlbaum.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 4-30). Hillsdale, NJ: Erlbaum.

Holdich, C.E., & Chung, P.W.H. (2003). A 'computer tutor' to assist children develop their narrative writing skills: Conferencing with HARRY. *International Journal of Human-Computer Studies, 59*, 631-669.

Kamler, B. (1980). One child, one teacher, one class: The story of one piece of writing. *Language Arts, 57*, 680-693.

MacDonald, N. H., Frase, L. T., Gingrich P. S., and Keenan, S.A. 1982. The Writer's Workbench: Computer Aids for Text Analysis. *IEEE Transactions on Communications 30(1)*, 105-110.

Perkins, D. A. (1985). The fingertip effect: How information-processing technology shapes thinking. *Educational Researcher, 14*, 11-17.

Scardamalia, M., & Bereiter, C. (1986). Research on written composition. In C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 778-803). NY: Macmillan.

Zellermayer, M., Salomon, G., Globerson, T., & Givon, H. (1991). Enhancing writing-related metacognitions through a computerized writing partner. *American Educational Research Journal, 28*, 373-391.