The Impact of Automated Essay Scoring on Writing Outcomes

Mark D. Shermis

Cynthia Wilson Garvan

Yanbo Diao

University of Florida

Paper presented at the Annual Meetings of the National Council on Measurement in Education, New York, NY, March 25-27, 2008.

Abstract

This study was an expanded replication of an earlier endeavor (Shermis, Burstein, & Bliss, 2004) to document the writing outcomes associated with automated essay scoring. The focus of the current study was on determining whether exposure to multiple writing prompts facilitated writing production variables (*Essay Score*, *Essay Length*, and *Number of Unique Words*) and decreased writing errors (*Grammar*, *Usage*, *Mechanics*, *Style*, *Organization & Development*) over time. The impacts of these variables were examined in analyses of 11,685 essays written by 2,017 students at four grade levels (grades 6-8, 10). The essays, written in response to seven different prompts, were scored by automated essay scoring. The results showed significant differences across the four grades and over time for each of the eight outcome variables. Peak essay performance occurred with 8th graders who also displayed the highest reduction of both domain errors. Specific types of error reduction were differentially associated with grade level. The implications of the results for future research incorporating writing genre are discussed.

The Impact of Automated Essay Scoring on Writing Outcomes

Introduction

An earlier study (Shermis, Burstein, & Bliss, 2004) showed that tenth-grade students who used the results of automated essay scoring as feedback in writing instruction wrote more (about 60 words per essay), scored higher on writing prompts (about .2 on a five-point scale), and made fewer writing errors (especially mechanical errors) over a 20-week period than did students who did not have these scores available. Since the entire sample was from one grade level, the study had limited generalizability. Accordingly, the present study was conducted over four grade levels thereby providing the opportunity to identify any trends in developmental writing improvement. Uniquely, scoring and feedback were accomplished throughout the use of automated essay scoring devices.

The potential for automated essay scoring in enhancing writing outcomes, especially in the K-12 environment (Myers, 2003; Vantage Learning, 2003), rests upon its integration with pre-writing activities, providing non-judgmental feedback, and interactively engaging students in "discourse" (Burstein & Marcu, 2003). Current developments have increased the availability and cost effectiveness of the technology.

Purpose of the Study

Most of the research to date in AES has focused on the validity of the scoring models (Elliot, 2003; Keith, 1998, 2003), i.e., to determine whether the score provided by humans raters of essays is faithfully replicated by computers. The present study extends these earlier studies to examine more fully the effectiveness of the diagnostic feedback provided by AES, i.e., to determine whether writing "errors" identified by the computer and reported to the students are reduced and whether writing scores and writing productivity increase with repeated exposure to

AES prompts. Establishing the nature of these outcomes, as they are differentially related to writing variables across grades, provides a basis for understanding student receptivity to feedback provided by AES, consequently serving the more precise development of curriculum and instruction.

Role of Feedback on Improving Writing Scores

To effectively improve performance, formative feedback, if immediately and specifically applied, has been found to be effective for writing instruction (Beach & Friedrich, 2006). As such, it contributes to automaticity and fluency of subsequent performances on most tasks, including editing. Effective feedback, being informative, provides students with the bases for modifying or changing metacognitions related to improving their work through such processes as the selection of content to be emphasized or by the strategies they use for studying; both of which are aligned with targeted outcomes. Non-judgmental feedback is constructive. Its emotionally neutral affective qualities contribute to positive motivation affecting the acceptability of constructive feedback .The lack of feedback, or even positive feedback, on the other hand, may lead to metacognitions that no revisions are required (Somers, 1994, 1997).

Teachers often work under demanding time constraints when evaluating written work which can dramatically affect the quality of their ratings for feedback. It is not unusual to find an instructor attempting to evaluate and provide feedback on a number of writing assignments by 100 students or more per assignment. In a study of the effects of writing feedback, Yagelski (1995) found that 81.7% of the essays evaluated for a twelfth grade advanced writing class assignment provided feedback directed at surface and stylistic changes rather than on the more constructive aspects of needed changes in content, processing, or creativity. It is to be noted that specific detailed comments that provide explanations of feedback, including open-ended questions and use of face-to-face conferences to elaborate on comments (Bardine, Bardine, & Deegan, 2000) are more informative and therefore more effective in producing specific changes than are comments that are simply evaluative. Simply revising drafts without being informative has minimal effects on improvement of performance (in this case student writing) (Van Gelderen, 1997).

Unfortunately much of the feedback students receive on their writing products comes in the form of vague, *pro forma*, global, or inconsistent written comments (Smith, 1997; Straub, 1996). Non-informative marginal notes teachers sometimes use when grading papers, such as "awkward" or "tighten up" or even such general statements as "good," or "needs improvement" are not helpful (Smith, 1997). The most helpful comments are those (a) which make specific suggestions for *what is to be improved* and *how to make those improvements* (Beach & Friedrich, 2006; Ferris, 2003) and (b) which explain the reason(s) for a rating of good or bad (Beach, 1989).

Beyond teacher comments, there are three other types of feedback related to that provided by AES. They are discourse analysis, reader (teacher or peer) based feedback, and self assessment.

Discourse analysis (Burstein and Marcu, 2000) is a means of providing relatively non-judgmental feedback. Essentially, the rater, reader, or computer indicates an impression of the main thesis of the essay, what the supporting arguments are, what the conclusion might be, and so on. The writer, in turn, is expected to make a revision based on the analysis that is provided. If there is a mismatch between rater and writer expectations (which are subjective), the writer continues with revisions until a match occurs. When writer expectations (which are subjective) match the analysis of the rater, revisions are no longer made.

“Reader-based” feedback is similar to discourse analysis. The reader (or rater) provides the writer with a description of the processes the rater used in reading a draft—e.g., “I expected more description,” “I found something that was confusing,” “I had anticipated a change in plot.” The mechanism for producing change here is that the reader provides a section-by-section analysis of how the writing impressed him/her. This type of running dialogue is specific and can be productive resulting in good potential for improvement in writing from one draft to another (Elbow, 1981; Johnston, 1983).

Self-assessment of writing portfolios, and more recently electronic portfolios, require students to reflect on the qualities of their writing: why they thought a piece writing was good (or bad), what criteria were used to formulate that judgment, and what might be done in the future to improve the writing (Stiggins, 2007). By engaging in this dialogue, they take responsibility for mastering the writing criteria, diagnosing when writing is not on par with expectations, and generating ideas that may help improve writing. Within this framework, the teacher becomes a collaborator who serves as a sounding board for student-generated ideas.

Automated Essay Scoring

Depending on the platform employed, Automated Essay Scoring (AES) provides (a) numerical ratings on specific traits of interest (e.g., grammar, style, usage), (b) specific examples of bad writing, and/or (c) a discourse dialogue of what it interprets to be the intent of the writer. Proponents of the technology argue that by allowing the computer to provide this kind of feedback, the instructor can focus on assisting with the creative or content-related aspects of the essay (Myers, 2003).

Holistic versus Trait Ratings

Although it is unnecessary to incorporate trait ratings to provide a wide array of feedback on writing, the literature on this topic may provide some insight as to how effective feedback in specific domains is likely to be. Traits typically look at dimensions of writing that are thought to be important such as *content*, *creativity*, *style*, *mechanics*, and *organization* (and all their varieties). Perkins (1983) suggested that the advantage of trait ratings over holistic ratings as stemming from its "precise, detailed description of student's writing ability for a specific rhetorical task" (p. 600). The additional information provided in the detailed description can be useful for sorting and classification decisions (e.g., placement decisions; Hamp-Lyons, 1995). Moreover, the information can provide a formative mechanism on which writers can base changes for improved submissions. The criticisms of the approach include the multiple evaluation procedures that may be required for a given task and the time-consuming process involved in generation of scoring guidelines (Perkins, 1983). Moreover, trait ratings have not been found to be more advantageous sources of feedback than holistic scores (Shermis, Koch, Page, Keith, & Harrington, 2002).

Achieving Reliable and Valid Scores

Automated Essay Scoring (AES) is the evaluation of written work via computers. Initial research restricted AES to English; it has recently been extended to other languages as well (Kawate-Mierzejewska, 2003, March; Vantage Learning, 2001, 2002). Most packages place documents within an electronic portfolio. They provide a holistic assessment of the writing which can be supplemented by trait scores based on an established rubric, and may provide qualitative critiques through discourse analysis. Most use ratings from humans as the criterion for determining accuracy of performance, though some of the packages will permit validation against other sources of information (e.g., large informational databases).

Scoring specific writing elements by AES.

Obviously, computers don't "understand" written prose in the same way that humans do, a point that may be unnerving until one reflects on ways alternative technologies achieve similar results. Thus, one can estimate the length of a wall using a traditional tape measure or employ a laser-pointing device to achieve similar results. The computer scores essays according to models of what human raters consider desirable and undesirable writing elements. Collections of these elements are referred to as "traits," the intrinsic characteristics of writing called "trins" (Page & Peterson, (1995). The specific elements are called proxies or "proxes" (Page & Petersen, 1995). The differentiation of "trins" and "proxes" is parallel to that of "latent" and "observed" variables in the social sciences: thus, the score on an IQ test might be thought of as a "prox" (specific element) for the underlying characteristics of the "trin" (conceptualization) intelligence.

AES software packages include computer programs that parse the essay text, for the purpose of identifying hundreds of prox variables ranging from simple to complex. A deceptively simple variable is essay length. Although raters value this attribute, the relationship to good writing is not linear but rather logarithmic; raters value the amount of writing output up to a point, but then they look for other salient aspects of writing once the quantity threshold is met. Similarly, the number of occurrences of "because" is a relevant feature. Although seemingly a superficial feature, it importantly serves as a proxy for the beginning of a dependent clause. And this, in turn, is reflective of sentence complexity.

Establishing a criterion for performance.

When human raters comprise the criterion against which rating performance is judged, AES engines work off of a statistical model developed using the following procedures: (a) Obtain a sample of (500) essays with (4-8) human ratings on each essay; (b) Randomly select (300) essays and regress the human ratings against the variable set available from various computational analyses of a text; (c) use a subset of consolidated feature variables, or the factor structure underlying a set of feature variables, in order to formulate a regression equation. The equation doesn't have to have a linear basis, but linear models are easier to explain; (d) cross-validate the regression equation on the 200 remaining essays to determine if the original regression line has suffered from shrinkage (Shermis, Burstein, & Leacock, 2006).

Reliability of AES evaluations.

Most of the evidence suggests that AES evaluations are equivalent to or higher than evaluations of reliability with human raters (Elliot, 2003; Landauer, Laham, & Foltz, 2003). All AES engines have obtained exact agreements with humans in the mid 80's and adjacent agreements in the mid-high 90's--slightly higher than the agreement coefficients for trained human raters. The slight edge for AES may be a function of the fact that the statistical models are based on more raters than one would typically find in a rating enterprise. Several validity studies have suggested that AES engines tap the same construct as that being evaluated by human raters. Page, Keith, & LaVoie (1995) examined the construct validity of AES, Keith (2003) summarized several discriminant and true score validity studies of the technology, and Attali & Burstein (2006) demonstrated the relationship between AES and instructional activities associated with writing.

AES is not without its detractors. Ericcson & Haswell (2006) performed a comprehensive critique of the technology from the perspective of those who teach post-secondary writing. Objections to the technology ranged from a concern about the ethics of using computers rather than humans to teach writing to the lack of synchronicity between how human graders approach the rating task and the process by which AES evaluates a writing sample to failed implementations of AES in university placement testing programs. Nevertheless, the positive contributions of AES far outweigh the negative. It is an increasingly pervasive assessment technology that is used for both assessment and instruction.

Method

Participants

The data for the present study were drawn from K = 13,091 essays contained in one of the standardization samples in the ETS *Criterion*^{SM} database. (*Criterion* is the instructional and portfolio component of a system that incorporates *e-rater*™ as an automated essay scoring component. *Criterion* and *e-rater* are described in more detail below.) The essays were written by students in grades 6, 7, 8, and 10 and were solicited from a cluster of 480 K–12 Educational Testing Service clients (districts or schools) across the United States comprising a pool of 160,000 users. No demographic information, other than grade distribution, was obtained. Even though a few students wrote up to 17 essays, the sample size dropped precipitously after 7 essays. Accordingly, the sample was restricted to those who participated in first seven writing assignments at their grade level. As a consequence the number of essays was reduced from N = 13,091 to K = 11,685. The student, grade, and essay distributions are shown in Table 1.

Instruments

*Criterion*SM is a web-based service developed by ETS for evaluating writing skills, instantaneous reports of scores, and diagnostic feedback. [For a detailed description of the system, see Burstein, Chodorow, & Leacock (2003)]. *Criterion* incorporates two complementary applications based on Natural Language Processing methods. One 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 composed 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 in *Critique* are used to identify five main types of grammar usage, and mechanical errors including agreement errors, verb formation errors, wrong word use, missing punctuation, and typographical errors. The detection of grammatical violations is corpus based and statistical.

The construction of e-rater version 2.0 models is given in detail in Attali and Burstein (2006). It is composed of 12 features used by e-rater v2.0[1] to score essays. The 12 features are associated with six areas of analysis: errors in grammar, usage, and mechanics (Leacock & Chodorow, 2003); style (Burstein, 2003); identification of organizational segments, such as thesis statement (Burstein et al., 2003); and vocabulary content (Attali & Burstein, 2006).

Eleven of the individual features reflect essential characteristics in essay writing and are aligned with human scoring criteria. The first six of the 11 features are contained in the *Critique* writing analysis tools, and reflect the kinds of feedback that human raters provide, though not necessarily in the same statistical form (Attali, 2004). These features include: (1) proportion

[1] As of this writing, the current version of e-rater is 3.0.

squared of grammar errors, (2) proportion of word usage errors, (3) proportion of mechanical errors, (4) proportion of style comments, (5) number of required discourse elements, (6) average length of discourse elements, (7) score assigned to essays with similar vocabulary, (8) similarity of vocabulary to essays with score of “6”, (9) number word types divided by number of word tokens, (10) log frequency of least common words, (11) average length of words, and (12) total number of words (Attali & Burstein, 2006).

Once the values of all 12 features are determined, *e-rater* uses them to score essays in a process that includes finding the weights of its features, determining appropriate scaling parameters, and assigning scores (Attali & Burstein, 2004, 2006).

The weights of individual features can be determined by simply applying a multiple linear regression technique with the standardized human-based score as an outcome and standardized feature scores as predictors. However, the weights of individual features can also be determined by content experts or by setting them to values determined during prior similar assessments. Attali and Burstein (2006) found that judgment-based weights are not less efficient than statistically obtained optimal weights (found through regression analysis). With *e-rater*, it is also possible to combine optimal and judgment-based weights of features. Generally, once essays’ *e-rater* continuous scores are determined, they are transformed to a set of ordinal essay ratings.

In addition to providing information used in formulating a predicted score for each essay, *e-rater* identifies and counts the number of errors each writer makes in five broad areas: grammar, usage, mechanics, style, and organization and development. Some of this information is reported both quantitatively and qualitatively to the writer in the form of feedback through the *Critique* program.

Procedure

Prompts were administered to students during the time period August 2005 to July 2006. There was no control over the order in which prompts were given and in some cases teachers were permitted to create their own prompts. There was no control over the time interval between prompts. As mentioned above, the prompts appeared in the *Criterion* electronic portfolio as a writing assignment and students had one hour to complete their work. In some cases, students could submit their work multiple times for evaluation. For the purposes of this study, only data from the last attempt was recorded. Students received both quantitative and qualitative feedback. Holistic scores on the essay were provided which ranged from 0 to 6 and the *Critique* program highlighted writing problems or provided a narrative about how the computer interpreted a particular aspect of writing.

Results

The variables of concern were grade level, essay order, the three production variables (*Essay Score*, *Essay Length*, *Number of Unique Words*) and the 47 error codes. Five error variables were derived by summing over items within each essay as follows: *Grammar* (9 items, E101- E109), *Usage* (7 items, E201- E207), *Mechanics* (11 items, E301- E311), *Style* (6 items, E401- E406), and *Organization* (9 items, E501- E3509). Summary statistics were computed and the data graphically displayed to identify outliers and/or impossible or implausible values, to summarize the data, and to check for distributional forms. There were extreme values in all error variables and in the *Number of Unique Words* variable. These variables were winsorized by replacing extreme values with the value of the 99th percentile.

Table 1 shows the means and standard deviations of the three production variables (*Essay Score*, *Essay Length*, *Number of Unique Words*) and the 47 error codes across all four grade levels. On average, students received an essay score of 4.04 (*SD* = 1.39), produced essays with310.29 words (*SD* = 156.93) long, and used 36.63 (*SD* = 9.97) unique words in the construction of their responses. The range of the error means runs from 0 (*E206: Preposition Errors* were never flagged) to 25.16 (*SD* =19.91) for *E401: Repetition of Words*. Most of the errors averaged less than one per essay, but a few had noteworthy distributions, including *E503: Supporting Ideas* (*M* = 11.66, *SD* = 8.25), *E507: Transitional Words and Phrases* (*M* = 4.83, *SD* = 3.77), and *E301: Spelling* (*M* = 4.04, *SD* = 6.05).

Figures 1-8 illustrate the trends over the three essay production variables and the five error domains by grade level for the duration of the seven writing assignments. Because the number of errors a writer makes may be influenced by the amount of writing generated in the essay, we controlled for essay length by creating a ratio of errors/number-of-words in the analyses summarized in Figures 4-8 and in all subsequent analyses. A set of figures showing error rates for all the 47 individual error codes is assembled in Appendix I.

A generalized linear mixed model (GLMM) was used to characterize grade effects and subject-specific effects over time (i.e., essay order). Longitudinal data methods (Verbeke & Molenberghs, 2000) allow for the correlation of within-subject measures (over time) and allow for mechanisms to incorporate missing data (e.g., missing at random, missing completely at random). Statistical analyses to address hypothesis was based on the following general linear mixed model:

$$Y_{ijk} = \mu + \alpha_i + d_{ij} + (\alpha\tau)_{\iota\kappa} + e_{ijk}$$

where,

μ, α_i , $(\alpha\tau)_{\iota\kappa}$,are fixed parameters

d_{ij} is the random effect associated with the j^{th} subject in group *i*

e_{ijk} is the random error associated with the j^{th} subject in group *i* at sequence time *k* with α_i testing for intercept, and $(\alpha\tau)_{\iota\kappa}$ testing for linear effect or slope. Autoregressive and unstructured variance-covariance matrices were considered. Competing models (with each variance-covariance structure) were run with parameters estimated by the maximum-likelihood estimation method. Akaike Information Criterion (AIC) was lower (indicating better fit) for models with unstructured variance-covariance matrices. The longitudinal models were then run using restricted maximum likelihood and unstructured variance-covariance matrices. Winsorized data were used for the longitudinal modeling.

Table 5 shows the results of the overall analysis based on the SAS Proc Mixed analysis routine with an assumed unstructured variance-covariance matix. The table is organized around the three production variables crossed with the five error domains. Note that for the eight outcomes there was a significant difference by *grade level*. With regard to the production variables, Scheffe post hoc comparisons across all grades were significant for *Essay Score*. In this sample, the trend of mean scores was a linear increase which peaked at grade eight and then dropped slightly at grade ten. The trend for *Essay Length* was linear; word production increased as grade level increased. Pairwise Scheffe post hoc differences across all the means are significant with the exception of eighth and tenth grade comparisons. Finally, the trend for *Number of Unique Words* parallels that of overall word production. Pairwise Scheffe post hoc differences across all the means are significant with the exception of eighth and tenth grade comparisons. As noted above, we controlled error domain vectors to account for the increasing length of the essays.

The table also displays a significant difference over time (*essay_order*grade*) over all eight outcomes. The regression estimates listed in Tables 6-13 show the directionality of the changes over time. Thus, three of the four regression estimates (*essay_order*grade*) for the

Essay Score variable are significantly positive. Similarly, with both *Essay Length* and *Number of Unique Words*, two of the four regression estimates are significantly positive. In the error scores of *Grammar* and *Mechanics*, all four regression lines were significantly negative, three of the four regression lines for *Usage* were signficantly negative, one of the regression lines for *Style* was significantly negative, and, one regression estimate for *Organization and Development* was significantly negative (tenth grade) while two were significantly positive (sixth & seventh grade).

To investigate which specific error codes changed significantly over time within each error type, difference scores were formed using the error code in the first essay and the corresponding error code in the last essay completed. Wilcoxon signed rank tests were used to test the null hypothesis of no change in median error. Table 14 shows the result of this analysis.

Discussion

This study essentially addressed three basic questions: Do productivity and error patterns for AES differ by grade level? Do the patterns change over repeated exposure to AES? Are particular types of error reduced due to the type and amount of feedback provided by AES?

It is probably not too surprising to have seen significant differences by grade on overall writing production, as one would assume that those in higher grades would perform better. Moreover, other studies have shown writing outcome differences by grade (Attali, 2006). However, in this sample, the pattern of peak *Essay Score* performance occurred at eighth grade rather than tenth grade . The differences in production variables (*Essay Length* and *Number of Unique Words*) by grade were significant, with an asymptote at eighth grade and tenth grade grade performance. This may be a maximum performance or it may be due to the fact that these were volunteer classrooms with unique (unknown) characteristics contributing to the observed performance outcomes.

There were significant differences across time by grade. The pattern of regression coefficients were, by and large, in the expected direction. For the production variables, the regression estimates were generally significantly positive and for the error domains, they were generally significantly negative. The follow-up analysis attempted to isolate particular variables within an error domain cluster to determine which variables might have been most sensitive to change over time holding constant *Essay Length*. Most critics of automated essay scoring (Ericsson & Haswell, 2006) suggest that if the technology were to be effective, it would be in the areas of pointing out grammatical, usage or stylistic areas—akin to what a sophisticated word processing package might do. However, in our analysis, the errors most flagged as significantly reduced over time were in *Mechanics* and *Organization and Development*.

In the follow-up analysis, we were also struck by several observations. First, while the error domains of *Mechanics* and *Organization and Development* had a number of signficantly reduced error codes over time, the patterns were not consistent from grade to grade This may in part be a function of (a) the types of prompts for which essays were written at each grade level, (b) instructional emphases at the different grades, or (c) differential developmental contribution to writing. Second,eighth graders triggered significant differences on 23 of the 47 error codes after just 7 essays; sixth-graders, 9;seventh-graders, 4; and tenth graders, 6. It is doubtful that automated essay scoring is more or less appropriate for a particular grade level, but some grade match between grade-levels and the portfolios in which the AES scoring engines are housed may have been appropriate. This is a question that is worth pursuing in future research endeavors.

Third, the sign test of most flagged significant error codes was negative which means that the flagged error code were significantly reduced. However, a few of them, primarily in *Organization and Development*, were positive. What this means is that students were making more errors at their last essay than they were at their first essay. Essentially this domain seems to be operating differently than the others, even though there was a statistical control for *Essay Length*. Despite the control, these variables can be inextricably confounded since the longer essays have the potential of containing more ideas *and* subsequent number of errors. It may be worth noting that for sixth graders, seven of the flagged error codes were significantly negative and two significantly positive; all four flagged error codes for seventh graders were significantly negative; eighth graders had 20 flagged error codes that were significantly negative and three that were significantly positive; and tenth 1 graders had four flagged codes that were significantly negative and two that were significantly positive.

Though the analysis showed good fit for a linear model, we were curious to explore whether a better fit might be had by using a quadratic model in the form:

$$Y_{ijk} = \mu + \alpha_i + d_{ij} + (\alpha\tau)_{\iota\kappa} + (\alpha\tau^2)_{\iota\kappa} + e_{ijk}$$

where,

μ, α_i , $(\alpha\tau)_{\iota\kappa}$,$(\alpha\tau^2)_{\iota\kappa}$ are fixed parameters
d_{ij} is the random effect associated with the j^{th} subject in group *i*
e_{ijk} is the random error associated with the j^{th} subject in group *i* at sequence time *k*

with α_i testing for intercept, $(\alpha\tau)_{\iota\kappa}$ testing for linear effect or slope, and $(\alpha\tau^2)_{\iota\kappa}$ assessing quadratic fit. Table 15 shows the result of this analysis, and in every case, the quadratic term is significant which means that this term accounts for additional variance in the prediction equation.

Attali (2006) suggested the use of grade level and writing genre as two vehicles for establishing teacher-generated prompts constructed “on-the-fly.” That is, it is likely that essays written to certain prompts will lead to more improvement than to other prompts. And, this effect is probably related to the sophistication of the writer (i. e., grade level). The results presented here do provide additional evidence for the use of grades as a norming dimension, but note that writing genre was an influential variable since the required data regarding prompt genre were unavailable for analysis. However, a future study might be able to link writing genre with specific error reduction outcomes. This information would be helpful to future researchers trying to hone in on the types of error reduction (or identifying likely productivity increases) contingent on the type of writing in which students are engaged.

It is no more reasonable to expect that all errors across all domains would be reduced with automated essay scoring than with any other scoring scheme. However, linking the error codes with specific genre and dimensions in the use of AES may in the long run provide more consistent expectations regarding the favorable effects of feedback on writing ability such as that demonstrated in this study showing significant differences over time in as few as seven essays. Such consistency in findings would be of inestimable value for those planning curicular instructional interventions to improve writing ability.

References

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

Attali, Y. (2006, April). *On-the-fly automated essay scoring.* Paper presented at the National Council on Measurement in Education, San Francisco, CA.

Attali, Y., & Burstein, J. (2004). *Automated essay scoring with e-rater V.2.0.* Paper presented at the Annual Meeting of the International Association for Educational Assessment, Philadelphia, PA.

Attali, Y., & Burstein, J. (2006). Automated essay scoring With e-rater V.2. *Journal of Technology, Learning, and Assessment*, *4*(3), Available from http://www.jtla.org.

Bardine, B., Bardine, M., & Deegan, E. (2000). Beyond the red pen: Clarifying our role in the response process. *English Journal, 90*(1), 94-101.

Beach, R. (1989). Showing students how to assess: Conferences. In C. Anson (Ed.), *Writing and response: Theory, practice, and research* (pp. 127-148). Urbana, IL: National Council of Teachers of English.

Beach, R., & Friedrich, T. (2006). Response to writing. In C. A. McArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of Writing Research* (pp. 222-234). New York, NY: Guilford Press.

Burstein, J. (2003). The E-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113-122). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Burstein, J., Chodorow, M., & Leacock, C. (2003). *Criterion: Online essay evaluation: An application for automated evaluation of test-taker essays.* Paper presented at the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence, Acapulco, Mexico.

Burstein, J., & Marcu, D. (2003). A machine learning approach for identification of thesis and conclusion statements in student essays. *Computers and the Humanities*, *37*(4), 455-467.

Elbow, P. (1981). *Writing with power* (2 nd ed.).

Elliot, S. (2003). Intellimetric: From here to validity. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71-86). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Ferris, D. R. (2003). *Response to student writing: Implications for second language students*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly*, *29*(759-762).

Johnston, B. (1983). Assessing writing.

Kawate-Mierzejewska, M. (2003, March, March 23). *E-rater software.* Paper presented at the Japanese Association for Language Teaching, Tokyo, Japan.

Keith, T. Z. (1998). *Construct validity of PEG.* Paper presented at the American Educational Research Association, San Diego, CA.

Keith, T. Z. (2003). Validity and automated essay scoring systems. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 147-168). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Leacock, C., & Chodorow, M. (2003). C-rater: Scoring of short-answer questions. *Computers and the Humanities, 37*(4), 389-405.

Myers, M. (2003). What can computers contribute to a K-12 writing program? In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary approach* (pp. 3-20). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Page, E. B., Keith, T., & Lavoie, M. J. (1995, August). *Construct validity in the computer grading of essays.* Paper presented at the annual meeting of the American Psychological Association, New York, NY.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76*(7), 561-565.

Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly, 17*, 651-671.

Shermis, M. D., Burstein, J., & Bliss, L. (2004, April). *The impact of automated essay scoring on high stakes writing assessments.* Paper presented at the annual meetings of the National Council on Measurement in Education, San Diego, CA.

Shermis, M. D., Burstein, J., & Leacock, C. (2006). Applications of computers in assessment and analysis of writing. In C. A. MacArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of Writing Research* (pp. 403-416). New York, NY: Guilford Publications.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement, 62*(1), 5-18.

Smith, S. (1997). The genre of the end comment: Conventions in teacher responses to student writing. *Collge Composition and Communication, 48*(2), 249-268.

Somers, N. (1994). Revision strategies of student writers and experienced adult writers. *College Composition and Rhetoric, 44*, 378-387.

Somers, N. (1997). Responding to student wriing. *College Composition and Rhetoric, 45*(2), 148-156.

Stiggins, R. (2007). *An introduction to student-involved assessment for learning* (5th ed.). Portland, OR: Assessment Training Institute.

Straub, R. (1996). The concept of control in teacher response: Defining the varieties of "directive" and "facilitative" commentary. *College Composition and Communication, 47*(2), 223-251.

Van Gelderen, A. (1997). Elementary students' skills in revising: Integrating quantitative and qualitative analysis. *Written Communication, 14*(3), 360-397.

Vantage Learning. (2001). *A Preliminary study of the efficacy of IntelliMetric™ for use in scoring Hebrew assessments*. Newtown, PA: Vantage Learningo. Document Number)

Vantage Learning. (2002). *A study of IntelliMetric™ scoring for responses written in Bahasa Malay* (No. RB-735). Newtown, PA: Vantage Learningo. Document Number)

Vantage Learning. (2003). *A true score study of grade 11 student writing responses using IntelliMetric™ Version 9.0* (No. RB-786). Newtown, PA: Vantage Learningo. Document Number)

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York, NY: Springer-Verlag.

Yagelski, R. (1995). The role of classroom context in the revision strategies of student writers. *Research in the Teaching of English, 29,* 216-338.

Authors Note

We wish to thank the Educational Testing Service for providing the data used in this study. The opinions expressed in this paper reflect those of the authors and not necessarily those of the Educational Testing Service. The study could not have been completed without the kind assistance of Jill Burstein at the Educational Testing Service, Alexandr Shneyderman of Miami Dade Public Schools, Bob Dedrick of the University of South Florida, and Frank DiVesta, Professor Emeritus of Penn State University. Electronic correspondence concerning this paper should be directed to mshermis@ufl.edu .

Table 1.

Sample Grade and Essay Distribution.

Grade	*N*	Percent	Number of Essays	Percent
6	402	19.9	1,815	15.5
7	356	17.7	1,913	16.4
8	721	35.7	4,560	39.0
10	538	26.7	3,397	29.1
Total	2,017	100.0	11,685	100.0

Table 2.

Types of Errors Identified by e-rater (Used by Permission).

Category	**Code Number**	**Element**
Grammar	101	Fragments
	102	Run-on sentences
	103	Garbled sentences
	104	Subject-verb agreement
	105	Ill-formed verbs
	106	Pronoun error
	107	Possessive error
	108	Wrong or Missing Word
	109	Proofread This!
Usage	201	Wrong article
	202	Missing or extra article
	203	Confused words
	204	Wrong form of word
	205	Faulty comparisons
	206	Preposition error
	207	Nonstandard verb or word form
Mechanics	301	Spelling
	302	Capitalize Proper Nouns
	303	Missing initial capital letter in a sentence
	304	Missing question mark
	305	Missing final punctuation
	306	Missing apostrophe

Table 2 (continued).

Types of Errors Identified by e-rater (Used by Permission).

Category	**Code Number**	**Element**
	307	Missing comma
	308	Hyphen error
	309	Fused words
	310	Compound words
	311	Duplicates
Style	401	Repetition of words
	402	Inappropriate words or phrases
	403	Sentences beginning with coordinating conjunctions
	404	Too many short sentences
	405	Too many long sentences
	406	Passive voice
Organization & Development	501	Thesis statement
	502	Main ideas
	503	Supporting ideas
	504	Conclusion
	505	Introductory material
	506	Other
	507	Transitional words and phrases
	508	Repetition of ideas
	509	Topic relationship and technical quality

Table 3.

Example of a Few Tenth Grade Prompts (Used by Permission)

PET CARE LETTER (workplace writing) Suppose you've just gotten a pet or an animal that you never owned before. Write a letter to a local pet store, pet owners' association, or veterinarian asking for information about how to care for your pet.
LONGER SCHOOL YEAR (writing for assessment) Some educators believe that students lose valuable learning time during the long summer vacation. They have proposed that students go to school all year round with shorter breaks during the year. What is your reaction to this proposal? Write a letter to your school board stating your position with reasons to support your point of view.
EFFECTIVE WORLD LEADER (writing for assessment) Select an American president or world leader who has governed most effectively. Write an essay in which you give reasons and examples to support your choice.
GLOBAL ISSUE (persuasive essay) Think about a global issue--achieving world peace or eliminating hunger and poverty--on which you can take a stand. Write a persuasive essay in which you support your position with good reasons and examples.
WRITE A REVIEW (response to literature) Think about a novel that you have read recently. Write a review for your school newspaper that explains the most interesting aspect of the book, such as its character, theme, setting, or plot.
LOCAL ISSUE (problem/solution) Think of a problem that people face in your neighborhood or school. Write an editorial to your local newspaper presenting a solution to the problem you have identified.
ENFORCING DRESS CODE (persuasive) High schools, restaurants, work places, and the military all use dress codes. Think about the reasons for instituting dress codes and why they might be enforced in each case. Then, select one example of the use of dress codes. Write an essay in which you argue the benefits or drawbacks of a dress code in that situation.

Table 4.

Means and Standard Deviations for the Production and Error Variables across all Grades

Variable	*N*	Minimum	Maximum	*Mean*	*SD*
Essay Score	11685	0	6	4.04	1.388
Essay Length	11685	10	2565	310.29	156.934
Number of Unique Words	11685	3	85	36.63	9.966
E101	11685	0	44	.53	1.288
E102	11685	0	6	.04	.270
E103	11685	0	3	.01	.130
E104	11685	0	27	.21	.695
E105	11685	0	10	.14	.461
E106	11685	0	2	.00	.067
E107	11685	0	15	.19	.564
E108	11685	0	7	.02	.177
E109	11685	0	9	.13	.410
E201	11685	0	4	.05	.256
E202	11685	0	16	.82	1.252
E203	11685	0	16	.70	1.326
E204	11685	0	3	.00	.053
E205	11685	0	3	.01	.087
E206	11685	0	0	.00	.000
E207	11685	0	5	.01	.151
E301	11685	0	133	4.04	6.047
E302	11685	0	43	.44	1.837
E303	11685	0	44	.33	1.257
E304	11685	0	6	.04	.216
E305	11685	0	12	.08	.413
E306	11685	0	39	.20	.835
E307	11685	0	5	.09	.338
E308	11685	0	9	.05	.312
E309	11685	0	10	.08	.399
E310	11685	0	8	.12	.425
E311	11685	0	3	.07	.275
E401	11685	0	424	25.16	19.912
E402	11685	0	3	.00	.074
E403	11685	0	15	.08	.670
E404	11685	0	102	1.53	4.059
E405	11685	0	4	.05	.284
E406	11685	0	6	.10	.376
E501	11685	0	15	1.53	1.479
E502	11685	0	44	2.14	1.470
E503	11685	0	207	11.68	8.253
E504	11685	0	20	2.49	2.171
E505	11685	0	35	2.11	2.495
E506	11685	0	25	.43	1.026
E507	11685	0	43	4.83	3.770
E508	11685	0	1	.03	.169
E509	11685	0	5	.02	.178

Table 5.

Fixed Effects Estimates for Grade across the Production Variables and Error Variable Clusters

Dependent Variable(s)	Effect	DF Numerator	DF Denominator	*F*	Pr > *F*
Essay Score	Grade	4	2017	4608.69	<.0001
	Essay Order * Grade	4	2017	32.72	<.0001
Essay Length	Grade	4	2017	2576.85	<.0001
	Essay Order * Grade	4	2017	13.48	<.0001
Number Unique Words	Grade	4	2017	8043.12	<.0001
	Essay Order * Grade	4	2017	18.38	<.0001
Grammar	Grade	4	2017	24.83	<.0001
	Essay Order * Grade	4	2017	3.45	0.0081
Usage	Grade	4	2017	18.71	<.0001
	Essay Order * Grade	4	2017	3.07	.0155
Mechanics	Grade	4	2017	59.45	<.0001
	Essay Order * Grade	4	2017	10.01	<.0001
Style	Grade	4	2017	152.69	<.0001
	Essay Order * Grade	4	2017	6.38	<.0001
Organization & Development	Grade	4	2017	21.96	<.0001
	Essay Order * Grade	4	2017	14.86	<.0001

Table 6.

Fixed Effects Regression Estimates on Essay Score by Grade for the Seven Essays

Effect	Grade	Estimate	Standard Error	*DF*	*t*-value	Pr > |*t*|
Grade	10	3.8364	0.0538	2017	71.26	<.0001
Grade	6	2.8555	0.0648	2017	44.00	<.0001
Grade	7	3.7719	0.0681	2017	55.32	<.0001
Grade	8	4.2814	0.0466	2017	91.74	<.0001
Essay_order*grade	10	0.0259	0.0097	2017	2.65	0.0080
Essay_order*grade	6	0.1131	0.0146	2017	7.74	<.0001
Essay_order*grade	7	0.0239	0.0145	2017	1.64	0.1004
Essay_order*grade	8	0.0668	0.0085	2017	7.83	<.0001

Table 7.

Fixed Effects Regression Estimates on Essay Length by Grade for the Seven Essays

Effect	Grade	Estimate	Standard Error	*DF*	*t*-value	Pr > |*t*|
Grade	10	329.25	5.6592	2017	58.18	<.0001
Grade	6	208.87	6.8009	2017	30.71	<.0001
Grade	7	315.34	7.0826	2017	44.52	<.0001
Grade	8	311.44	4.9008	2017	63.55	<.0001
Essay_order*grade	10	0.1640	0.9997	2017	0.16	0.8697
Essay_order*grade	6	7.1261	1.5183	2017	4.69	<.0001
Essay_order*grade	7	-1.9052	1.4555	2017	-1.31	0.1907
Essay_order*grade	8	4.7653	0.8681	2017	5.49	<.0001

Table 8.

Fixed Effects Regression Estimates on Number of Unique Words by Grade for the Seven Essays

Effect	Grade	Estimate	Standard Error	*DF*	*t*-value	Pr > |*t*|
Grade	10	38.4790	0.3825	2017	100.59	<.0001
Grade	6	28.1168	0.4602	2017	61.10	<.0001
Grade	7	35.9105	0.4801	2017	74.79	<.0001
Grade	8	37.5375	0.3313	2017	113.32	<.0001
Essay_order*grade	10	0.0002	0.0689	2017	0.00	0.9968
Essay_order*grade	6	0.7633	0.1047	2017	7.29	<.0001
Essay_order*grade	7	0.0032	0.1008	2017	0.03	0.9744
Essay_order*grade	8	0.2700	0.0598	2017	4.51	<.0001

Table 9.

Fixed Effects Regression Estimates on Grammar by Grade for the Seven Essays

Effect	Grade	Estimate	Standard Error	*DF*	*t*-value	Pr > |*t*|
Grade	10	0.4819	0.1040	2017	4.63	<.0001
Grade	6	0.9707	0.0991	2017	9.79	<.0001
Grade	7	0.5501	0.1146	2017	4.80	<.0001
Grade	8	0.3858	0.0969	2017	3.98	<.0001
Essay_order*grade	10	-0.0525	0.0237	2017	-2.21	0.0272
Essay_order*grade	6	-0.0721	0.0248	2017	-2.91	0.0037
Essay_order*grade	7	-0.0620	0.0268	2017	-2.31	0.0210
Essay_order*grade	8	-0.0739	0.0226	2017	-3.27	0.0011

Table 10.

Fixed Effects Regression Estimates on Usage by Grade for the Seven Essays

Effect	Grade	Estimate	Standard Error	*DF*	*t*-value	Pr > |*t*|
Grade	10	0.7003	0.1158	2017	6.05	<.0001
Grade	6	0.8312	0.1090	2017	7.62	<.0001
Grade	7	0.6669	0.1271	2017	5.25	<.0001
Grade	8	0.3978	0.1081	2017	3.68	0.0002
Essay_order*grade	10	-0.0524	0.0264	2017	-1.99	0.0470
Essay_order*grade	6	-0.0558	0.0272	2017	-2.05	0.0403
Essay_order*grade	7	-0.0440	0.0300	2017	-1.47	0.1421
Essay_order*grade	8	-0.0854	0.0252	2017	-3.38	0.0007

Table 11.

Fixed Effects Regression Estimates on Mechanics by Grade for the Seven Essays

Effect	Grade	Estimate	Standard Error	*DF*	*t*-value	Pr > |*t*|
Grade	10	1.9312	0.4363	2017	4.43	<.0001
Grade	6	6.1818	0.4199	2017	14.72	<.0001
Grade	7	2.9558	0.4842	2017	6.10	<.0001
Grade	8	1.0222	0.4044	2017	2.53	0.0116
Essay_order*grade	10	-0.4318	0.0931	2017	-4.64	<.0001
Essay_order*grade	6	-0.5122	0.0960	2017	-5.33	<.0001
Essay_order*grade	7	-0.2961	0.1055	2017	-2.81	0.0050
Essay_order*grade	8	-0.4589	0.0887	2017	-5.17	<.0001

Table 12.

Fixed Effects Regression Estimates on Style by Grade for the Seven Essays

Effect	Grade	Estimate	Standard Error	*DF*	*t*-value	Pr > |*t*|
Grade	10	14.0273	1.1705	2017	11.98	<.0001
Grade	6	25.1711	1.1037	2017	22.81	<.0001
Grade	7	18.3853	1.2860	2017	14.30	<.0001
Grade	8	18.6858	1.0910	2017	17.13	<.0001
Essay_order*grade	10	0.07607	0.2678	2017	0.28	0.7764
Essay_order*grade	6	-0.2013	0.2738	2017	-0.74	0.4622
Essay_order*grade	7	0.4510	0.3030	2017	1.49	0.1368
Essay_order*grade	8	-0.7262	0.2559	2017	-2.84	0.0046

Table 13.

Fixed Effects Regression Estimates on Organization and Development by Grade for the Seven Essays

Effect	Grade	Estimate	Standard Error	*DF*	*t*-value	Pr > |*t*|
Grade	10	1.1102	0.4758	2017	2.33	0.0197
Grade	6	2.1714	0.4820	2017	4.50	<.0001
Grade	7	-1.2197	0.5352	2017	-2.28	0.0228
Grade	8	3.0822	0.4321	2017	7.13	<.0001
Essay_order*grade	10	-0.2670	0.0838	2017	-3.18	0.0015
Essay_order*grade	6	0.3673	0.0877	2017	4.18	<.0001
Essay_order*grade	7	0.1948	0.0952	2017	2.05	0.0409
Essay_order*grade	8	-0.1389	0.0796	2017	-1.74	0.0814

Table 14.

Error Codes with Significant Differences (Wilcoxon Sign Test) between Essay 1 and Last Essay Completed.

Grade	Cluster	Error Code	N	*S*	Pr > |*S*|
10	Usage	202	511	-5,482.5	.0194
10	Mechanics	301	511	-7,262	.0086
10	Mechanics	302	511	-1,359.5	.0212
10	Style	406	511	1,365	.0013
10	Organization & Development	501	511	7,774.5	.0114
10	Organization & Development	507	511	-9,394.5	.0041
6	Grammar	105	379	-514.5	.0026
6	Mechanics	301	379	-7242	.0001
6	Mechanics	303	379	-1,856.5	<.0001
6	Mechanics	305	379	-340.5	.0455
6	Style	401	379	-10,170.5	<.0001
6	Organization & Development	501	379	-6393	.0002
6	Organization & Development	503	379	-4,709	.0147
6	Organization & Development	504	379	3,476.5	.0008
6	Organization & Development	506	379	2,917.5	<.0001
7	Grammar	104	333	-813.5	.0077
7	Mechanics	307	333	-500.5	.0002
7	Style	401	333	-3992	.0038
7	Organization & Development	502	333	-6,156	<.0001
8	Grammar	101	695	-3664	.0423
8	Grammar	102	695	-620	.0001
8	Grammar	103	695	-46.5	.0432
8	Grammar	107	695	-2,841	.0001
8	Grammar	108	695	-78	.0452
8	Grammar	109	695	-946.5	.0085
8	Usage	202	695	-8,627	.0010
8	Usage	203	695	-9,568	<.0001
8	Mechanics	301	695	-32,662.5	<.0001
8	Mechanics	302	695	-3,226.5	<.0001
8	Mechanics	303	695	-1,565.5	.0004
8	Mechanics	306	695	-1,744.5	<.0001
8	Mechanics	307	695	-644.5	.0008
8	Mechanics	308	695	-200	.0272
8	Mechanics	309	695	-419	.0005
8	Mechanics	310	695	-729	.0248
8	Mechanics	311	695	-275.5	.0447
8	Style	401	695	-22,490	<.0001
8	Organization & Development	501	695	-23,489	<.0001
8	Organization & Development	502	695	10,977.5	.0317
8	Organization & Development	504	695	21,461.5	<.0001
8	Organization & Development	505	695	18,597	<.0001
8	Organization & Development	509	695	-181.5	.0067

Table 15.

Fixed Effects Estimates for Grade across the Production Variables and Error Variable Clusters using a Quadratic Model

Dependent Variable(s)	Effect	DF Numerator	DF Denominator	*F*	Pr > *F*
Essay Score	Grade	4	2017	1359.34	<.0001
	Essay Order*Grade	4	2017	10.08	<.0001
	Essay Order^2 * Grade	4	2017	12.79	<.0001
Word Length	Grade	4	2017	52669.80	<.0001
	Essay Order*Grade	4	2017	7.58	<.0059
	Essay Order^2 * Grade	4	2017	5.12	<.0014
Number Unique Words	Grade	4	2017	6142.97	<.0001
	Essay Order*Grade	4	2017	10.54	<.0012
	Essay Order^2 * Grade	4	2017	4.21	<.0055
Grammar	Grade	4	2017	17.66	<.0001
	Essay Order*Grade	4	2017	6.65	<.0001
	Essay Order^2 * Grade	4	2017	4.14	0.0024

Table 14 (Continued).

Dependent Variable(s)	Effect	DF Numerator	DF Denominator	F	Pr > F
Usage	Grade	4	2017	17.83	<.0001
	Essay Order*Grade	4	2017	4.73	0.0008
	Essay Order^2 * Grade	4	2017	3.24	0.0117
Mechanics	Grade	4	2017	45.72	<.0001
	Essay Order*Grade	4	2017	22.45	<.0001
	Essay Order^2 * Grade	4	2017	16.21	<.0001
Style	Grade	4	2017	98.63	<.0001
	Essay Order*Grade	4	2017	7.33	<.0001
	Essay Order^2 * Grade	4	2017	6.09	<.0001
Organization & Development	Grade	4	2017	14.33	<.0001
	Essay Order*Grade	4	2017	0.38	0.8199
	Essay Order^2 * Grade	4	2017	1.31	0.2641

Figure 1.

Trend for Essay Scores Across Four Grade-Levels after Seven Essays.

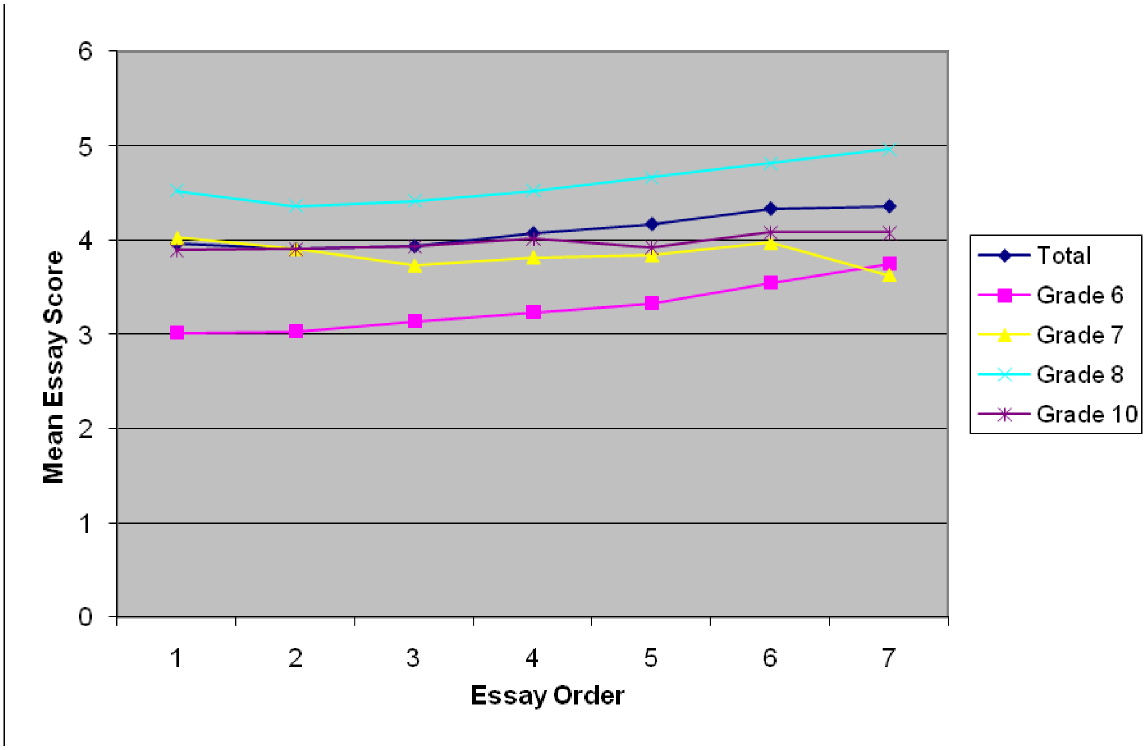

Figure 2.

Trend for Essay Length Across Four Grade-Levels after Seven Essays.

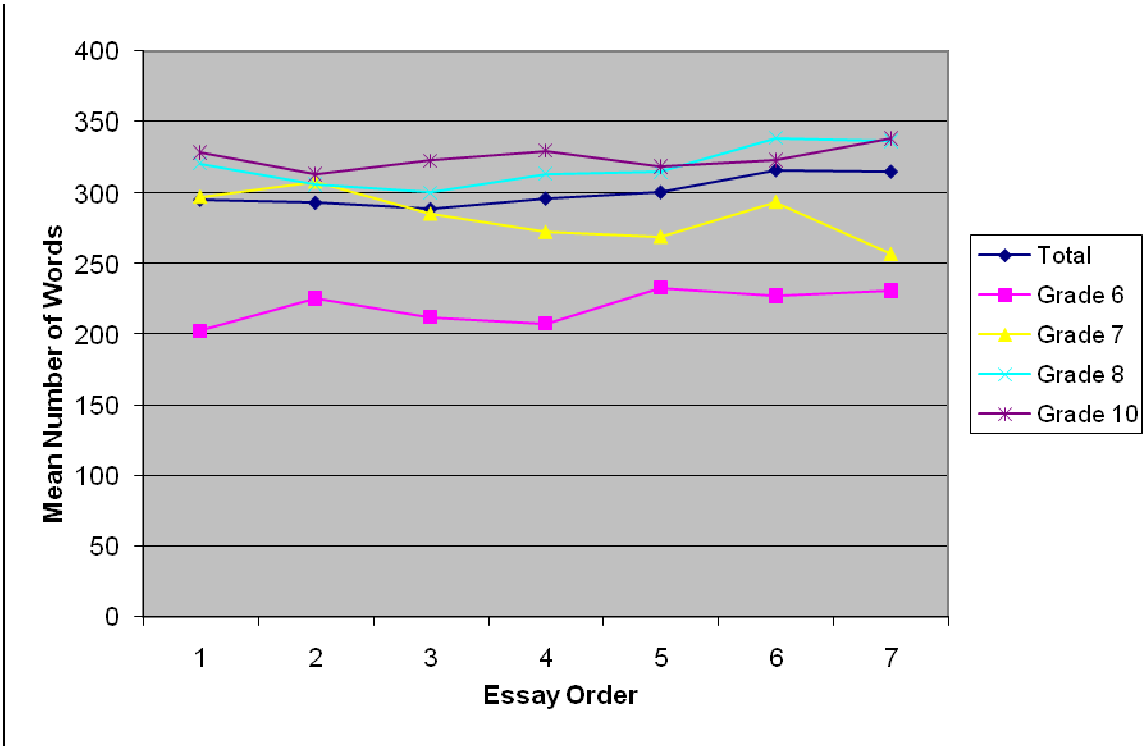

Figure 3.

Trend for Number of Unique Words Across Four Grade-Levels after Seven Essays.

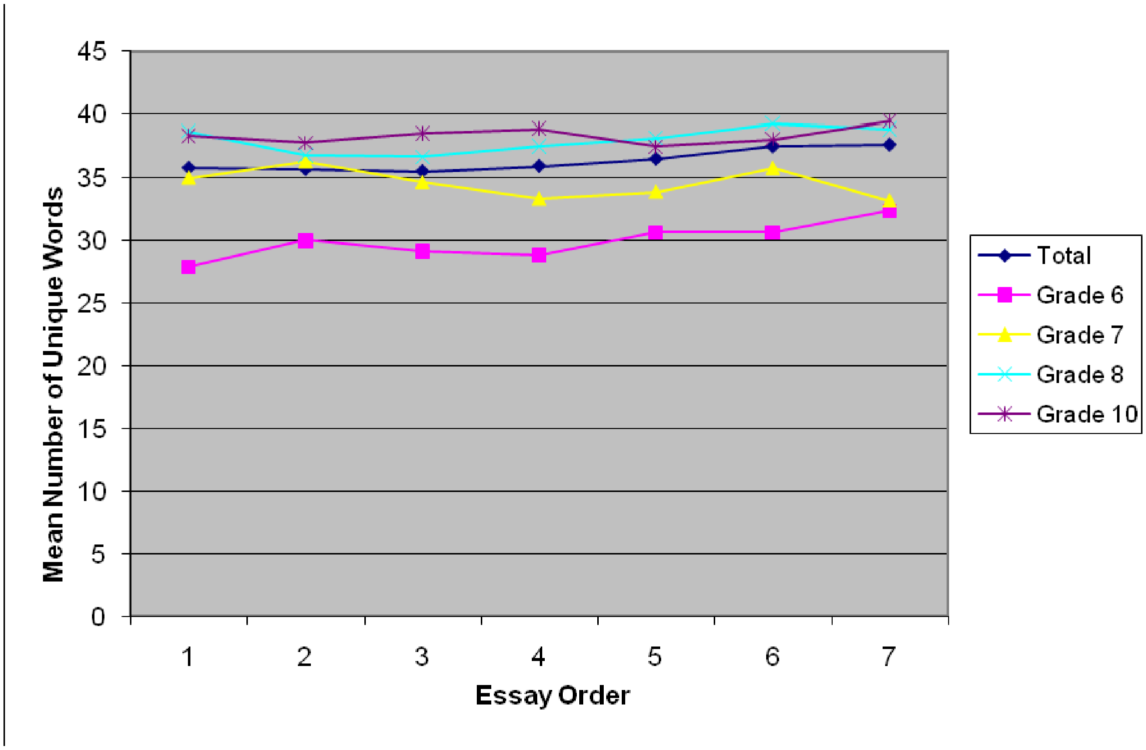

Figure 4.

Trend for Grammar Error Across Four Grade-Levels after Seven Essays. Note: Means Adjusted for Number of Words Written.

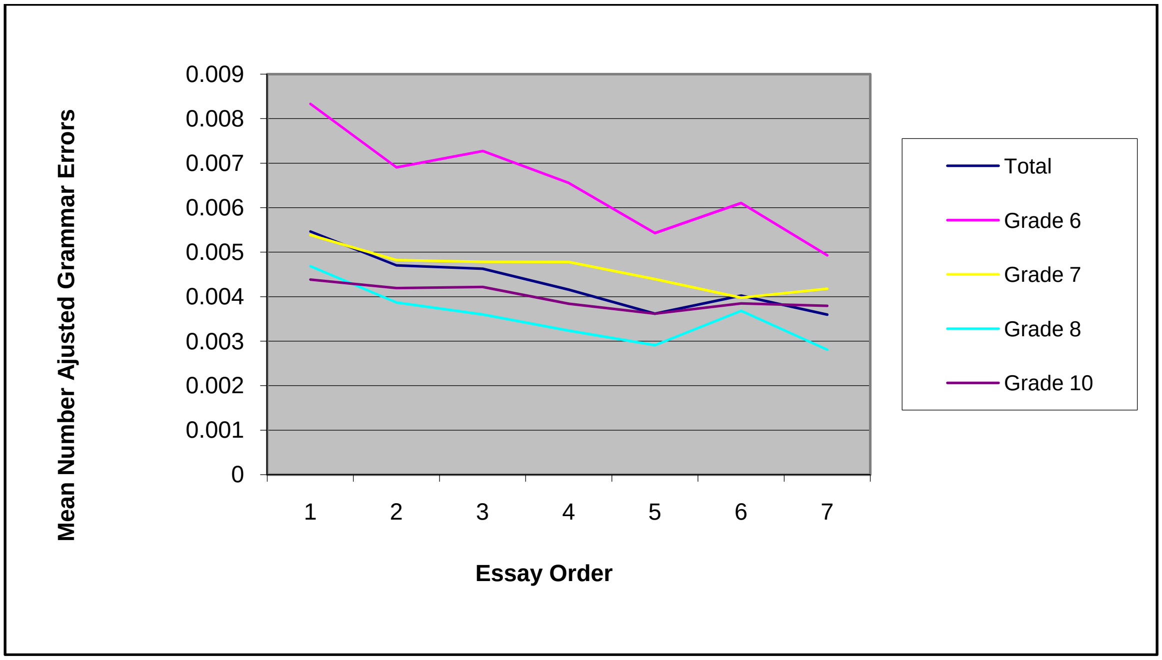

Figure 5.

Trend for Usage Error Across Four Grade-Levels after Seven Essays. Note: Means Adjusted for Number of Words Written.

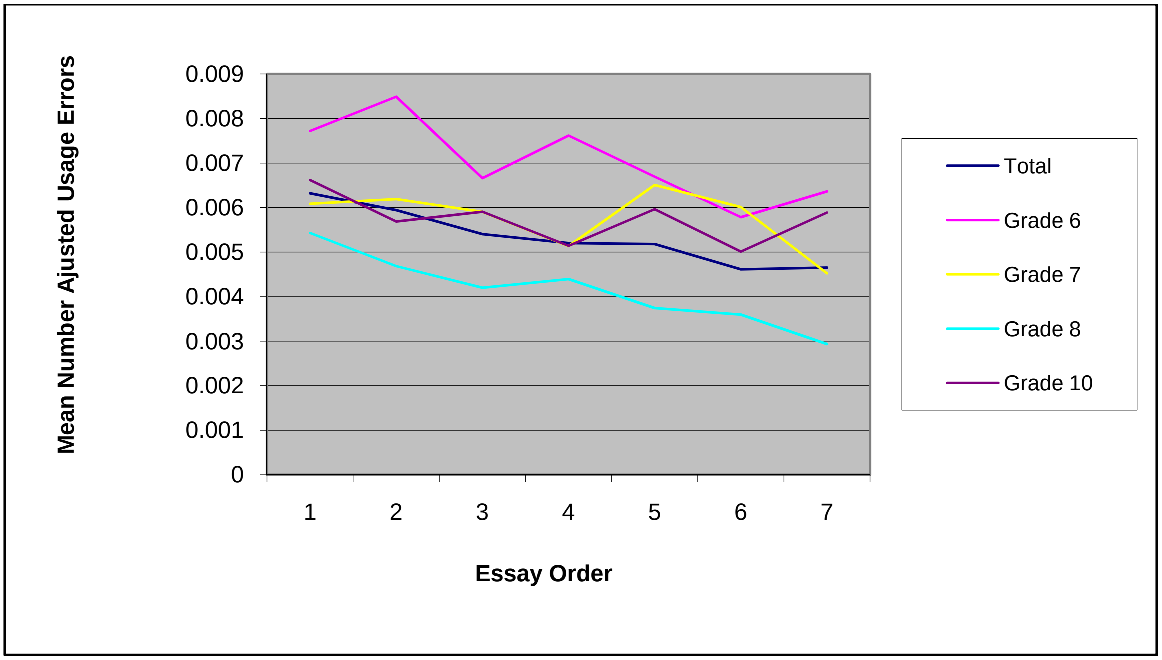

Figure 6.

Trend for Mechanics Error Across Four Grade-Levels after Seven Essays. Note: Means Adjusted for Number of Words Written.

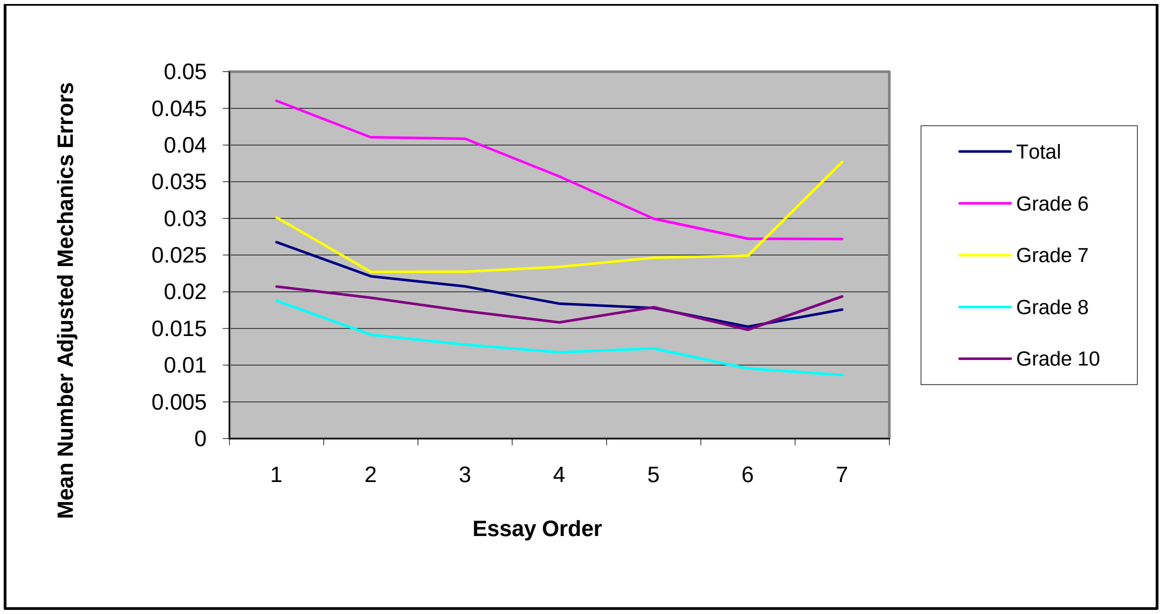

Figure 7.

Trend for Style Errors Across Four Grade-Levels after Seven Essays. Note: Means Adjusted for Number of Words Written.

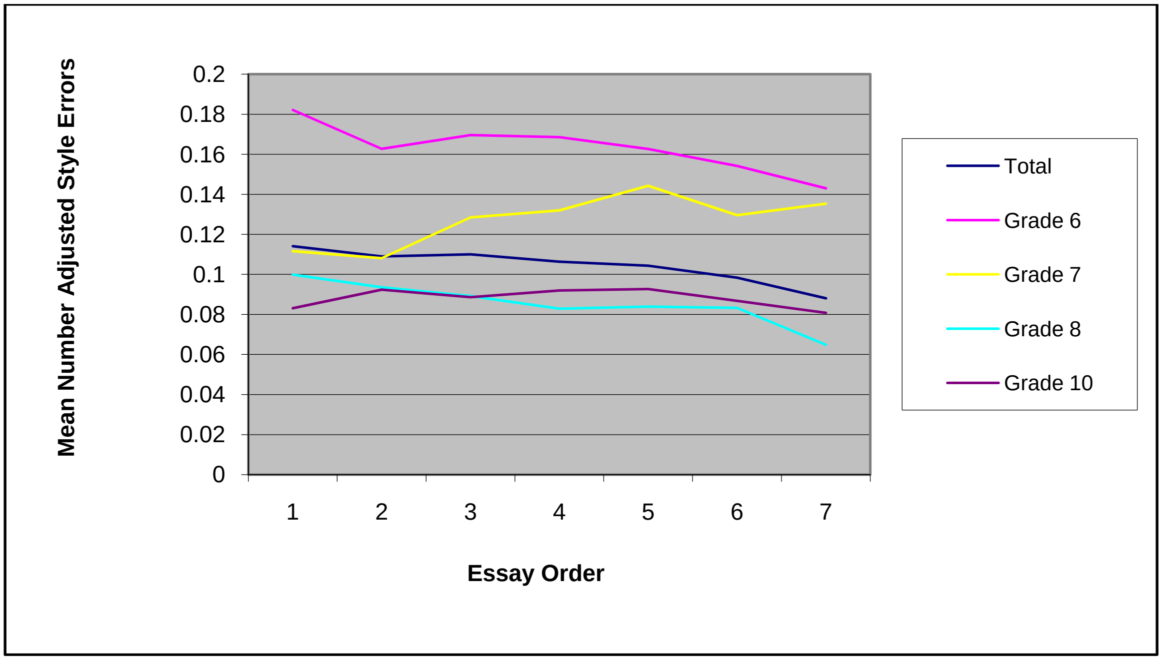

Figure 8.

Trend for Organization and Development Errors Across Four Grade-Levels after Seven Essays.

Note: Means Adjusted for Number of Words Written.