Automated Assessment of Paragraph Quality: Introduction, Body, and Conclusion Paragraphs

Rod D. Roscoe[1], Scott A. Crossley[2], Jennifer L. Weston[1], Danielle S. McNamara[1]

[1]Department of Psychology, Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152
[2]Department of Applied Linguistics/ESL, Georgia State University, Atlanta, GA 30302
rdroscoe@memphis.edu, sacrossley@gmail.com, jlweston@memphis.edu, dsmcnamr@memphis.edu

Abstract

Natural language processing and statistical methods were used to identify linguistic features associated with the quality of student-generated paragraphs. Linguistic features were assessed using Coh-Metrix. The resulting computational models demonstrated small to medium effect sizes for predicting paragraph quality: introduction quality r^2 = .25, body quality r^2 = .10, and conclusion quality r^2 = .11. Although the variance explained was somewhat low, the linguistic features identified were consistent with the rhetorical goals of paragraph types. Avenues for bolstering this approach by considering individual writing styles and techniques are considered.

Writing Practice and Assessment

Effective writing is a critical skill related to academic and professional success (Geiser & Studley, 2001; Kellogg & Raulerson, 2007), yet large-scale assessments often show that writing proficiency is elusive for many students (National Commission on Writing, NCW, 2003).

Strategy instruction, writing practice, and individualized feedback are needed to improve students' writing skills (Graham & Perin, 2007; Kellogg & Raulerson, 2007). Students must be taught strategies for enacting the writing process – prewriting, drafting, and revision – along with the knowledge needed to employ the strategies. Students must also have opportunities to practice these developing strategies and receive timely and individualized feedback throughout the learning process. Practice and feedback are key for students to reflect on their writing and understand how their use of writing strategies impacts writing quality.

Although highly effective, strategy instruction with ample practice and feedback requires significant time and effort. Classroom instructors are constrained in their ability to give personal and detailed feedback on student writing as a result of available instructional time, increasing class sizes, and a focus on standardized tests (NCW, 2003).

Automated Essay Scoring

Automated essay scoring (AES) – the use of computers to grade student essays – allows students to practice writing and receive feedback, without adding to teachers' burdens (Dikli, 2006). Writing can be assessed via combinations of statistical modeling, natural language processing (NLP), Latent Semantic Analysis (LSA), artificial intelligence (AI) and machine learning, and other methods.

Systems such as e-rater (Burstein, Chodorow, & Leacock, 2004) and IntelliMetric (Rudner, Garcia, & Welch, 2006) rely primarily on NLP and AI. First, a corpus of essays is annotated to identify target essay elements (e.g., topic sentences). Essays are then automatically analyzed along many linguistic dimensions, and statistical analyses extract features that discriminate between higher and lower-quality essays. Finally, weighted statistical models combine the extracted linguistic properties into algorithms that assign grades to student essays.

The Intelligent Essay Assessor (IEA, Landauer, Laham, & Foltz, 2003) uses LSA to assess essays. LSA assumes that word meanings are often determined by their co-occurrence with other words. Texts are represented in a word-by-context matrix. Context refers to sentences, paragraphs, or whole texts. Singular value decomposition reduces the number of dimensions to capture semantic structure. Using LSA, student essays are compared to a benchmark corpus of pre-scored essays to assess semantic similarity. Essay scores are based on the overlap between student essays and the benchmarks. LSA does not require annotation, model-building, human ratings, or syntactic parsing; essentially, the benchmark corpus is the model.

AES systems have successfully increased opportunities for student writing with feedback. Researchers also report positive correlations and high percent agreement with

human raters (Dikli, 2006). Two main objections to AES are that it lacks humanist sensitivity and detection is limited by available algorithms (Hearst, 2002). Automated essay scorers, and their reliance on statistical regularities, may not capture writers' style, voice, or other individual expressive differences. Thus, despite progress, automated scoring systems are still under development, with opportunities to expand in many areas.

Assessing Paragraph Quality

In this project, we contribute to AES research by assessing the quality of canonical components of the five-paragraph essay: introduction, body, and conclusion paragraphs (Albertson, 2007; Johnson, Smagorinksy, Thompson & Fry, 2007). In five-paragraph essays, students first state their thesis and arguments in an engaging introduction. Subsequently, each argument forms the topic sentence of a body paragraph, in which evidence is offered to support that claim. Finally, the author's thesis and claims are summarized in a conclusion paragraph that demonstrates the unity and significance of ideas.

Detractors have argued that the five-paragraph essay stifles creativity and leads to formulaic writing (Albertson, 2007; Dean, 2000). However, for new and struggling writers, the structure can provide an objective schema for organizing and communicating one's ideas. Moreover, for better or worse, the five-paragraph essay is an important aspect of standardized testing, such as the SAT Reasoning Test (SAT).

Prior research has sought to automatically detect introduction, body, and conclusion paragraph types by combining linguistic and LSA methods (Crossley, Dempsey, & McNamara, under review) using Coh-Metrix (Graesser et al., 2004; McNamara, Crossley, & McCarthy, 2010). In the Crossley et al. study, initial paragraphs (versus middle and final paragraphs) were shorter, contained less word overlap, and fewer positive logical connectives (e.g., *also, then*), and contained more specific, meaningful, and imageable words. The directness of these paragraphs, combined with evocative and meaningful word choices, was consistent with introduction paragraph goals: concisely stating one's position and arguments in a way that grabbed the reader's attention. Middle paragraphs were longer, contained more given information (maintained a common thread of ideas), and less imageable and familiar words. The greater length and consistency of these paragraphs might have been necessary for the development of evidence and examples to support a single, coherent topic sentence. In addition, the use of less imageable or familiar words may have indicated the authors' elaboration upon specific or abstract principles. Lastly, final paragraphs were shorter, and used words that were less meaningful and specific, but more familiar. Conclusions also displayed less given information, more content word overlap, and more positive logical connectives. These linguistic features were consistent with the rhetorical goal of providing a concise and accessible summary of one's position and arguments without adding new evidence or examples.

These paragraph features were used to develop a model capable of detecting paragraph type in a corpus of student writing. The reported model performed well above chance and the accuracy of the model (65%) was nearly identical to the accuracy of human judges (66%). Overall, these results suggested that meaningful properties of introduction, body, and conclusion paragraphs could be detected through automated assessment methods.

An important question is how such properties relate to paragraph quality. Crossley et al. (under review) reported post-hoc analyses showing an interaction between detection accuracy and quality. Paragraphs that were rated more highly by humans were easier to classify, by both humans and the model, than were poorly-rated paragraphs. Higher quality paragraphs may have been more likely to enact appropriate rhetorical forms (e.g., stating a clear thesis in the introduction), which aided detection.

The remainder of this paper reports linguistic analyses and the development of model to assess paragraph quality more directly. Our goal is to examine if there are linguistic properties that can be used to discriminate between well-written versus poorly-written introduction, body, or conclusion paragraphs.

Method

Paragraph Corpus and Scoring

We collected 201 essays written by 201 college freshmen at Mississippi State University (MSU). The essays were based on two SAT writing prompts that asked writers to discuss whether people should admire heroes or celebrities or whether originality is possible. The essays were timed (25 minutes) and no outside referencing was allowed. We selected essays with at least three paragraphs (N = 180), and labeled initial paragraphs as introduction paragraphs (n = 180), middle paragraphs as body paragraphs (n = 403), and final paragraphs as conclusion paragraphs (n = 180). Paragraphs were randomized within these groupings and separated into training and test sets using a 67/33 split.

Essays were rated using a SAT-like rubric modified by two experts in linguistics and cognitive psychology. The rubric comprised items 9 items: effective lead and clear purpose (introduction), topic sentences, transitions, organization, and unity (body), perspective and conviction (conclusion), and mechanics. Each item generated a score between 1 (minimum) and 6 (maximum), with equal distance between each step.

Six expert raters (with advanced degree in English and 3+ years of composition teaching experience) used the rubric to rate each essay. Raters initially scored a set of 20 essays, and a Pearson correlation was conducted on raters' responses. If correlations did not exceed r =. .50 (p < .05) on all items, the ratings were reexamined. After reaching an inter-rater reliability of r = .70, each rater independently evaluated the 180 essays in the corpus. Once final ratings

were collected, differences between raters were calculated. If the difference in ratings on an item was less than 2, an average score was computed. If the difference was greater than 2, raters had the opportunity to discuss and revise their evaluation. All correlations between raters after adjudication were greater than .60.

An exploratory factor analysis was conducted to confirm that human scores on rubric items conformed to the intended rubric design. A Bartlett's test of sphericity was significant ($p < .001$), and the Kaiser-Meyer-Olin measure of sampling adequacy exceeded .90 for initial and subsequent analyses suggesting that the data was factorable. The scree plot suggested the extraction of four factors. Principal axis factoring using varimax rotation also identified four factors. All items loaded onto their respective factors with eigenvalues greater than .50, and these factors overlapped with expected subscales. The items that loaded onto the first factor (introductions) were *effective lead, clear purpose,* and *clear plan*. The items that loaded onto the second factor (conclusions) were *perspective* and *conviction*. The items that loaded on the third factor (body paragraphs) were *topic sentences, paragraph transitions, organization*, and *unity*. Finally, the last factor (mechanics) had only item: *mechanics*.

Factor-based scores for each paragraph were computed by averaging the item scores related to that paragraph type. These scores were used in a subsequent regression analysis using Coh-Metrix variables to identify potential linguistic differences accounting for paragraph quality.

Coh-Metrix Indices and Analyses

We examined numerous Coh-Metrix measures related to cohesion, lexical sophistication, syntactic complexity, and basic text features (e.g., length). For a more thorough overview of what is assessed by Coh-Metrix, see Graesser et al. (2004) or McNamara et al. (2010).

Cohesion. Cohesion is a key aspect of understanding language structure and how connections within in a text influence coherence and text comprehension (Kintsch & van Dijk, 1978). Cohesion increases when key words, arguments, and ideas overlap across sentences (coreference), in similar locations and contexts (minimal edit distance), and when new information is related to prior discourse (givenness). Cohesion is also improved by smoothly linking ideas (connectives), avoiding dense and abstract logical relations *(*logical operators), indicating causal relations (causal cohesion), using motion and spatial information to bolster text meaning (spatial cohesion), and conveying the temporal dynamics of the point itself (temporal cohesion).

Lexical sophistication. Lexical sophistication refers to the writer's use of advanced vocabulary and word choice to convey ideas. Lexical sophistication is a key element of human ratings of text quality (Crossley & McNamara, 2010; McNamara et al., 2010). Lexical sophistication is captured by assessing the type and amount of information provided by the words in a text. Words are assessed in terms of rarity (frequency), abstractness (concreteness), evocation of sensory images (imagability), salience (familiarity), and number of associations (meaningfulness). Words can also vary in the number of senses they contain (polysemy) or levels they have in a conceptual hierarchy (hypernymy). More broadly, texts may differ in the overall range of vocabulary employed (lexical diversity).

Syntactical complexity. The grammatical structure of the text is also an important indicator of human evaluations of text quality (McNamara et al., 2010). Difficult syntactic constructions (syntactic complexity) include the use of embedded constituents, and are often dense, ambiguous, or ungrammatical (Graesser et al., 2004). More uniform constructions (syntactic similarity) result in less complex syntax that is easier to process.

Analyses. To assess relationships between Coh-Metrix indices and paragraph quality, correlations were calculated between each measure and human ratings. Typically, many Coh-Metrix variables from each category demonstrate significant correlations. However, because many measures within a category tap overlapping constructs, they are often highly inter-correlated. Multicollinearity was addressed by selecting the 10 indices (based on a 20:1 ratio of texts to indices) with the highest correlation to ratings, but which did not display multicollinearity with other indices.

Selected indices were entered into a multiple regression to predict human paragraph quality ratings for the training set of paragraphs. The resulting *B* weights and constant terms from the training set regression were next used to estimate how well the model would function on an independent data set (the paragraphs held back for the test set). The model produced an estimated value for each paragraph in the test set, which was correlated with the subscale score to determine the strength of the model. A final analysis was conducted on the entire corpus of texts.

Results

Introduction Paragraphs

Table 1 presents the Coh-Metrix indices correlated with introduction quality. Introductions were rated more highly when they contained more given information (*LSA given/new*), displayed higher temporal (*tense repetition*) and causal (*causal particles and verbs*) cohesion, and used *positive logical connectors*. Introductions also received higher scores when writers showed a broader vocabulary (*lexical diversity MTLD*), using less common words (*content word frequency*) that were more specific (*noun hypernymy*). Higher quality introductions also employed more varied syntax (*mean syntax similarity*) and complexity *(number of words before main verb*).

For the training set, a linear regression (step-wise) was conducted including the 10 variables. Four variables were significant predictors: *number of words*, $t = 5.14$, $p < .001$; *content word frequency*, $t = -2.80$, $p < .01$; *ratio of causal particles and verbs*, $t = 2.31$, $p < .05$; and *tense repetition*, *t*

= 2.21, $p < .05$. The overall model was significant, $F_{4,115} = 13.14$, $p < .001$, $r = .56$, $r^2 = .31$, indicating that the combination of the four variables accounted for 31% of the variance in the introduction ratings (Table 2).

Coh-Metrix Variable	*r*
Overall Text	
Number of words	.42
Cohesion	
LSA given/new information	.21
Tense repetition	.16
Causal particles and verbs	.15
Positive logical connectors	.15
Lexical Sophistication	
Lexical diversity MTLD	.23
Content word frequency	-.22
Noun hypernymy	.16
Syntactic Complexity	
Mean syntax similarity	-.21
# of words before main verb	.17

Table 1. Correlations with introduction ratings.

The subsequent model for the test set yielded $r = .38$, $r^2 = .14$. The total set of paragraphs yielded $r = .50$, $r^2 = .25$, showing that the combination of variables accounted for 25% of the variance in all introduction paragraph ratings. Table 2 provides r^2 values from the stepwise regression and feature weights for the final model.

Variable	*r*	r^2	*ß*	*B*	*SE*
Number of words	.46	.21	.008	.404	.00
Word frequency	.50	.25	-.652	-.219	.23
Causal particles/verbs	.53	.28	.185	.181	.08
Tense repetition	.56	.31	.285	.172	.13

Table 2. Introduction paragraph regression analysis.

Results suggest several detectable properties of good introductions. Longer introductions may be more effective at conveying the writer's position and arguments, as opposed to short introductions that do not offer enough information to establish the writer's stance. The use of less frequent words may indicate a deeper vocabulary, and temporal cohesion improves readability. Finally, causal cohesion may indicate that the introduction previews clear arguments; that is, statements that claim causal relations between concepts or offer reasons to support a position.

Body Paragraphs

Table 3 presents the Coh-Metrix indices that correlated with body paragraph quality. Body paragraphs were rated more highly if they contained more given information (*LSA given/new*), words in diverse positions *(MED content word stems*), and spatial cohesion (*locational nouns*). Better body paragraphs tended to use a broader vocabulary (*lexical diversity MTLD*), including words that were less frequent and familiar, and more concrete, specific, and imageable.

Coh-Metrix Variable	*r*
Overall Text	
Number of words	.14
Cohesion	
MED content word stems	.19
Locational nouns	.12
LSA given/new information	.12
Lexical Sophistication	
Content word frequency	-.23
Content word familiarity	-.19
Lexical diversity MTLD	.14
Word hypernymy	.14
Content word concreteness	.12
Content word imagability	.11

Table 3. Correlations with body ratings.

A linear regression analysis (stepwise) was conducted including the ten variables from the training set. Four variables were significant predictors: *MED for content word stems*, $t = 3.74$, $p < .001$; *content word familiarity*, $t = -3.73$, $p < .001$; *locational nouns*, $t = 2.42$, $p < .05$; and *LSA givenness*, $t = 2.36$, $p < .05$. The overall regression model was significant, $F_{4,262} = 9.37$, $p < .001$, $r = .35$, $r^2 = .12$, indicating that the combination of the four variables accounted for 12% of the variance in the body ratings. Table 4 summarizes the data for the four variables.

The model for the test set yielded $r = .25$, $r^2 = .06$. The total set of paragraphs yielded $r = .32$, $r^2 = .10$, demonstrating that the combination of variables accounted for 10% of the variance all in body paragraph ratings. Table 4 provides r^2 values from the stepwise regression and feature weights for the final model.

Variable	*r*	r^2	*ß*	*B*	*SE*
MED content stems	.22	.05	1.451	.398	.22
Word familarity	.30	.09	-.019	.005	-.22
Locational nouns	.33	.11	.002	.001	.14
LSA given/new	.35	.12	1.501	.635	.14

Table 4. Body paragraph regression analysis.

Results suggest that better body paragraphs displayed a deeper vocabulary and more varied sentence structure. The use of locational nouns may indicate the writer's use of specific examples or cases as evidence, such as referring to events "at home" or "in school." A higher degree of givenness suggests that better body paragraphs maintained a common thread of ideas, rather than haphazardly jumping between disparate themes.

Conclusion Paragraphs

Correlations between Coh-Metrix indices and conclusion quality are provided in Table 5. Conclusion ratings were

associated with given information (*LSA given/new*), words in diverse positions *(MED all words*), and spatial cohesion (*location and motion words*). Better conclusions also maintained a consistent tense (*tense and aspect repetition*), with overlap among arguments *(argument overlap*). Lexically, conclusion paragraphs contained more uncommon words *(content word frequency*) and were conceptually more specific *(noun hypernymy*). The syntax appeared to be simpler, with sentences that were similarly constructed (*mean syntax similarity*) with fewer high-level constituents per word.

Coh-Metrix Variable	*r*
Overall Text	
Number of words	.23
Cohesion	
MED all words	.24
LSA given/new information	.22
Tense repetition	.21
Argument overlap	.18
Location and motion words	.16
Lexical Sophistication	
Content word frequency	-.21
Noun hypernymy	.18
Syntactic Complexity	
Mean syntax similarity	.22
High level constituents/word	-.18

Table 5. Correlations with conclusion ratings

For the training set of paragraphs, three variables were significant in a step-wise linear regression conducted using the above 10 variables: *number of higher level constituents per word*, $t = -3.43$, $p < .001$; *minimal edit distance for all words*, $t = -2.42$, $p < .05$; and *noun hypernymy*, $t = 2.24$, $p < .05$. The overall model was significant, $F_{3,116} = 8.95$, $p < .001$, $r = .43$, $r^2 = .19$, demonstrating that the combination of the three variables accounted for 19% of the variance in the paragraph ratings. Data are summarized in Table 6.

The model for the test set yielded $r = .14$, $r^2 = .02$. The total set of paragraphs yielded $r = .33$, $r^2 = .11$, showing that the combination of variables accounted for 11% of the variance in the ratings for all conclusion paragraphs. Table 6 provides r^2 values from the stepwise regression and feature weights for the final model.

Variable	*r*	r^2	*ß*	*B*	*SE*
Constituents per word	.33	.11	-2.901	-.293	.84
MED all words	.39	.15	.602	.203	.25
Noun hypernymy	.43	.19	.148	.192	.07

Table 6. Conclusion paragraph regression analysis.

Results suggest that better conclusions express more specific ideas using accessible, yet varied syntax. This pattern is consistent with the rhetorical goal of concluding an essay with a straightforward summary of one's ideas that provide the reader with a "big picture" understanding of the writer's position.

Discussion

The purpose of this study was to explore the extent to which linguistic features alone, ignoring other structural and semantic variables, were able to account for variance associated with the quality of particular types of paragraphs. Linguistic analyses of introduction, body, and conclusion paragraphs using Coh-Metrix revealed several properties associated with paragraph quality. Some features were common across all types: length, givenness of information, and vocabulary. Not surprisingly, paragraphs that were longer received higher ratings, perhaps because they contained more elaborated arguments or evidence. Better paragraphs also contained more given information, maintaining cohesion and comprehensibility of ideas. Lastly, several measures of lexical sophistication were predictive of paragraph quality, such as word frequency, hypernymy, and lexical diversity. Paragraphs received higher scores when the writers displayed a deeper and more varied choice of vocabulary. These results mimic those reported by McNamara et al. (2010) regarding the entire essays.

In this study, we found that linear regression models based on these indices were predictive, although not strongly, of human ratings of paragraph quality. These models also highlighted features of particular importance for each of the different paragraph types. For example, introductions seemed to benefit from greater causal cohesion, perhaps indicative of paragraphs that stated a clear thesis supported by relevant arguments. In contrast, givenness was especially important for body paragraphs. For example, evidence presented in body paragraph should relate to the topic sentence and build on prior statements. Finally, syntactic simplicity may have been important for conclusions. Perhaps the better conclusions were those that summarized main ideas straightforwardly, and offered an accessible take-home message.

One important question is why the regression models did not perform better. The models only predicted 25% of the variance in introduction ratings, 10% for body ratings, and 11% for conclusion ratings. Several limitations may account for these results.

One problem is the small corpus available. To ensure that the essays in our corpus could minimally contain an introduction, body, and conclusion paragraph, all essays with only 1 or 2 paragraphs were excluded. The resulting sample was much smaller than what may have been ideal.

Another concern rests with our assumptions about paragraph type. We labeled paragraphs in the initial position as "introductions," final paragraphs as "conclusions," and all others as "body" paragraphs. This categorization was probably not accurate in some cases. Some essays may have lacked an introduction (jumping straight into the evidence) or conclusion (running out of

time and ending with a body paragraph). Such cases would have added noise to the analyses.

In addition to sampling issues, it is also worthwhile to consider broader concerns about automated essay scoring. Specifically, the approach reported here may lack of sensitivity to individual writer characteristics such as "style." In one introduction paragraph in our sample, the writer began with leading questions, "What makes a hero? What makes a celebrity?" The use of leading questions is a common rhetorical technique for engaging the reader. However, there are many available techniques, such as anecdotes or establishing a controversy. In the introduction example below, the writer drew upon historical references:

> The United Stated of America was founded on the principle of innovation. From the very first landing at Plymouth to social media, America has set the bar high for innovation and invention. The United States Patent Office holds millions of ideas of unique products and methods. From these findings, people certainly can be original.

Writers' use of such techniques represents a possible source of individual variation or style that might be easily missed by AES methods based on statistical aggregates. Does use of such techniques improve paragraph or essay quality? If so, which techniques?

We are currently investigating this question and developing new key word measures, n-gram analyses, and word-classification techniques to detect various elements related to rhetorical strategies. For example, knowing that autobiographical anecdotes contain a high incidence of personal pronouns, first-person perspective, bigrams such as "When I" or "In my" and that historical anecdotes include proper nouns, references to the passage of time, and specific dates or times, allows us to develop indices to measure these introduction types. In addition, we are also developing new indices to identify topical adherence (e.g., by measure key word overlap or semantic co-referentiality with the prompt) along with n-gram analyses that identify rhetorical strategies (e.g., enumeration, persuasion, exemplification) and lexical items specific to paragraph type quality (e.g., the use of concluding statements in the final paragraph). By combining various approaches, we will increase our ability to detect whether and to what extent these techniques are employed. These methods will increase the range of essay features we are able to incorporate into models of writing quality, which should improve the accuracy of our automated paragraph scoring.

Acknowledgments

This research was supported in part by the Institute for Education Sciences (IES R305A080589). Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the IES. We would like to thank our expert raters: L. Bogard, B. Campbell, B. Hagenston, M. Kardos, A. Leonard, M. Price, and D. White.

References

Albertson, B. (2007). Organization and development features of grade 8 and grade 10 writers: A descriptive study of Delaware Student Testing Programs (DSTP) Essays. *Research in the Teaching of English, 41*, 435-464.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing system. *AI Magazine, 25*, 27-36.

Crossley, S., Dempsey, K., & McNamara, D. (Under review). Classifying paragraph types using linguistics features: Is paragraph positioning important? *Research in the Teaching of English*.

Crossley, S. A., & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. *Proceedings of the 32nd annual conference of the Cognitive Science Society*

Dean, D. (2000). Muddying boundaries: Mixing genres with five paragraphs, *English Journal, 90,* 53-56.

Dikli, S. (2007). An Overview of automated essay scoring of essays. *Journal of Technology, Learning, and Assessment*, 5(1). Retrieved Nov. 5, 2010 from www.jtla.org.

Geiser, S. & Studley, R. (2003). UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. Paper presented at the Meeting of the Board of Admissions and Relations with Schools of the University of California.

Graesser, A., McNamara, D., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, 36, 193-202.

Graham, S. & Perin, D. (2007). A Meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99,* 445-476.

Hearst, M. ed. (2002). The debate on automated essay grading. *IEEE Intelligent Systems*, *15*, 22-37

Johnson, T., Smagorinksy, P., Thompson, L., & Fry, P. (2003). Learning to teach the five-paragraph theme. *Research in the Teaching of English*, *38,* 136-176.

Kellogg, R. & Raulerson, B. (2007). Improving the writing skills of college students. *Psychonomic Bulletin and Review, 14*, 237-242.

Kintsch, W. & van Dijk, T. (1978). Towards a model of text comprehension and production. *Psychological Review*, *85*, 363-394.

Landauer, T., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education, 10,* 295-308.

McNamara, D., Crossley, S., & McCarthy, P. (2010). The linguistic features of quality writing. *Written Communication*, *27*, 57-86.

McNamara, D., Louwerse, M., McCarthy, P., & Graesser, A. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47,* 292-330.

National Commission on Writing. (2003). *The Neglected "R."* NY: College Entrance Examination Board.

Rudner, L., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric essay scoring system. *Journal of Technology, Learning, and Assessment*, 4(4). Retrieved Nov. 5, 2010 from www.jtla.org.