Internal Usability Testing of Automated Essay Feedback in an Intelligent Writing Tutor

[1]Rod D. Roscoe, [1]Laura K. Varner, [1]Zhiqiang Cai, [1]Jennifer L. Weston, [2]Scott A. Crossley, and [1]Danielle S. McNamara

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, lkallen1, zcai, jlweston, dsmcnamr}@memphis.edu, sacrossley@gmail.com

Abstract

Research on automated essay scoring (AES) indicates that computer-generated essay ratings are comparable to human ratings. However, despite investigations into the accuracy and reliability of AES scores, less attention has been paid to the feedback delivered to the students. This paper presents a method developers can use to quickly evaluate the usability of an automated feedback system prior to testing with students. Using this method, researchers evaluated the feedback provided by the Writing-Pal, an intelligent tutor for writing strategies. Lessons learned and potential for future research are discussed.

Introduction

The development of writing proficiency requires extended practice guided by individualized feedback (Kellogg & Raulerson, 2007). Sadly, teachers are often limited in their opportunities to provide feedback on student writing due to limited time and increasing class sizes (National Commission on Writing, 2003). One solution has been the use of automated essay scoring (AES). AES utilizes sophisticated software to evaluate the structure, content, and overall quality of written prose (Shermis & Burstein, 2003). By automating portions of the grading and feedback process, students are afforded more opportunities for writing practice, with fewer burdens placed on instructors.

AES research has primarily focused on assessing the accuracy of automated scores (Warschauer & Ware, 2006), but few studies address the construction or evaluation of the feedback given to students. This is an important omission because, although reliable algorithms are necessary for accurate scoring, students can only benefit if provided with clear and usable feedback (Shute, 2008). Ideally, usability data would be collected from authentic users. However, student essay corpora, along with expert human ratings, are costly and time-consuming to obtain. During the feedback development process, it is valuable to be able to test and refine feedback rapidly and cheaply. For these reasons, we developed an alternative "in-house" usability method that enables faster feedback system testing while requiring fewer resources. The method may prove beneficial in detecting potential student difficulties with the system feedback. This paper briefly reviews research on AES, describes the Writing-Pal intelligent tutoring system, and summarizes lessons learned from our internal testing of automated writing feedback.

Automated Essay Scoring

AES systems assess writing using diverse methods, including statistical modeling, natural language processing (NLP), Latent Semantic Analysis (LSA), and additional techniques from the field of artificial intelligence (AI) (Dikli, 2006; Shermis & Burstein, 2003).

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 corpus annotation, model-building, or syntactic parsing; essentially, the benchmark corpus is the model.

AES systems are not without objections (Hearst, 2002; Wang & Brown, 2007). Computers cannot "understand" the content of an essay in the same manner as a human. It is also possible to "trick" some grading systems with nonsensical texts that are grammatically, structurally, or thematically sound (Powers, Burstein, Chodorow, Fowles, & Kukich, 2001). However, available research finds that many systems are able to provide automated scores that are comparable to human scores (Warschauer & Ware, 2006).

On average, reported correlations between human and system scores range from .80 to .85 (Warschauer & Ware, 2006). For example, studies on IEA (Landauer et al., 2003) have found a mean correlation of .81 between human and system scores. Correlations were slightly higher for standardized test essays (.85) than essays written for classroom assignments (.73). Rudner et al. (2006) conducted two studies using IntelliMetric. The mean correlation between human and automated score was .83 in Study 1 and .84 in Study 2.

Percent agreement is reported in two ways: "perfect" (human and system score match exactly) and "perfect + adjacent" (human and system score are within 1 point of each other). Across two IntelliMetric studies, Rudner et al. (2006) reported perfect percent agreement ranging from 42% to 65%. Their perfect + adjacent agreement ranged from 92% to 100%. Attali and Burstein (2006) reported perfect agreement for one version of e-rater ranging from 46% to 58% (depending on grade level and test).

Another method for evaluating AES systems is to test whether students' essay quality improves. For example, Attali (2004) analyzed over 9000 essays (6^{th}-12^{th} grade) written by Criterion users over one year, comparing scores of original and final drafts. Final versions earned higher scores (effect size = .47), were longer, and contained fewer mechanical errors. Essays particularly improved in stating and supporting main ideas, and incorporating conclusions. Attali (2004) attributed these improvements to receiving feedback via Criterion, but over the course of one year, it is likely that maturation and teacher instruction also played roles in the improvement. No control condition (i.e., a non-Criterion comparison) was reported.

Other evaluations have taken place over a shorter time span, but with more experimental control. Rock (2007) analyzed essays written by nearly 1500 9^{th} grade students, half of whom worked with Criterion. The other half received only teacher feedback. Essays were evaluated holistically (based on a 6-level NAEP rubric) and in terms of mechanics. Mean holistic scores did not significantly differ by condition. Students who wrote and revised essays using Criterion did not produce better essays than students who received only teacher feedback. However, there was a small but significant difference in mechanics favoring the Criterion condition (effect size = .15).

Criterion efficacy was further examined by Kellogg et al. (2010) in a study that manipulated how much feedback was given. University students wrote and revised three essays over a 3-week period. Students either received feedback on all first drafts ("continuous feedback"), only one of their first drafts ("intermittent feedback"), or none of their drafts ("no feedback"). Results indicated that revised drafts received higher holistic scores (ranging from 1 to 6) from Criterion than original drafts, but there was no effect of condition. Essay quality improved regardless of whether automated feedback was provided. However, continuous feedback was more effective than no feedback in reducing grammar, mechanics, usage, and style errors.

In sum, available research on AES efficacy suggests that AES tools can help students improve their writing, most clearly for mechanics. However, Kellogg et al. (2010) raise an important point:

> Does automated feedback target the skills and abilities most appropriate to the development of students' writing skills? Probably not. A college instructor may be less concerned about errors in mechanics, and most interested in how a student can structure an essay to clearly convey ideas. Accordingly, feedback might address whether an essay has a logical thesis statement or whether main points have adequate support and elaboration.

In contrast to studies testing score reliability, very few publications evaluate the *usability of system feedback*. However, both are critical for the success and efficacy of automated scoring systems. Accurate evaluations are only useful to students if they are usable (Kluger & DeNisi, 1996; Shute, 2008). In this paper, we describe a method for evaluating feedback prior to system deployment. This method may help to troubleshoot and improve feedback during development, thus increasing the likelihood of eventual student performance gains. This testing was carried out within the Writing-Pal system (McNamara et al., in press).

The Writing-Pal and Formative Feedback

The Writing-Pal (W-Pal) is an intelligent tutoring system (ITS) for writing. Writing Strategy Modules cover freewriting, planning, introduction building, body building, conclusion building, paraphrasing, cohesion building, and revising. Each module consists of an interactive lesson that uses vicarious learning with animated agents (Craig, Driscoll, & Gholson 2004). Modules also include challenges that allow students to practice the strategies via game-like activities. Writing practice occurs in the Essay Writing Module, in which students author essays on SAT-style prompts. Students receive automated feedback after submitting their essays to W-Pal. Thus, although W-Pal is a tutoring system, it incorporates many AES elements.

W-Pal's scoring algorithms are under development (McNamara, Crossley, & McCarthy, 2010). Ultimately, W-Pal will combine many methods, including key word and N-gram analyses, NLP, and LSA. A key resource for this work is Coh-Metrix (McNamara, Louwerse, McCarthy, & Graesser, 2010), which assesses text features related to

cohesion, lexical sophistication, syntactic complexity, and more. At the time of this study, a simplified algorithm was used for pilot testing. More sophisticated algorithms could not be implemented due to time constraints, programming resources, and school access. Holistic scores, ranging from 1 (minimum) to 6 (maximum), were based on basic features such as text length, occurrence of non-words, and use of key words. This simplicity was a motivator of our desire to evaluate the usability of the feedback.

W-Pal provides *formative* and *scaffolded* feedback on student essays and writing strategies. Formative feedback reinforces the strategies taught in the lessons and uses reflective questions to remind students of important goals. The feedback also directs students toward lessons or challenges that contain further instruction.

W-Pal feedback is also scaffolded. For example, students who struggle to produce *any* text may not be ready to implement feedback about cohesion. Instead, these students may gain more from planning. This scaffolding is implemented as a series of threshold-based algorithms: 1) legitimacy (proportion of non-words), 2) length (number of words), 3) relevance (occurrence of key words), and 4) structure (number of paragraphs). Feedback is delivered for the highest level attained in the series of thresholds. Once essays pass these thresholds, they receive holistic feedback encouraging overall revision. Depending on the quality of individual sections, they also receive suggestions for introductions, bodies, and/or conclusions. These specific paragraphs types were also evaluated with respect to length and keywords. To prevent repetitiveness, five versions of every feedback response were developed.

The example below shows feedback for an essay that is "too short" and recommends ways to expand the essay:

> Effective writers put forth effort to make sure that the reader can understand the ideas presented. This essay might be expanded in several ways to communicate your ideas more completely.
>
> * One way to expand your essay is to add additional relevant examples and evidence
>
> * Another way to improve an essay is to provide more details that support your arguments
>
> * Have you created a flow chart or a writing road map to help you organize your ideas?
>
> * Trying using the Planning Lesson strategies to make sure your essay is not missing key information

Method

In this project, we evaluated the usability of W-Pal feedback by submitting previously collected student essays to W-Pal's grading algorithm. Two members of the development team (the first and second authors) then revised those essays based solely on the feedback delivered. Finally, the revised essays were resubmitted to W-Pal to assess improvement.

Essay Corpus

We collected 201 essays written by college students at Mississippi State University. The essays were based on two SAT-like writing prompts on "Originality" or "Heroes." The essays were timed (25 minutes) and no outside referencing was allowed. The essays were rated using a SAT scoring rubric. The rubric produced holistic ratings ranging from 1 (minimum) to 6 (maximum).

Six expert raters used the rubric to rate each essay. After reaching an inter-rater reliability of $r = .70$, pairs of raters independently evaluated the essays in the corpus. Once final ratings were collected, differences between raters were calculated. If the difference in ratings were 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.

Revision Corpus. Using stratified, random sampling, we selected three example essays from each prompt and each available scoring level. Very few essays scored a 1.0; therefore, we collapsed the 1.0 and 1.5 levels. The collapsed 1.0-1.5 level also contained more Originality essays ($n = 5$) than Heroes essays ($n = 1$). In addition, no essay scored above a 5.0. Thus, the final revision corpus contained 48 total essays across eight scoring levels.

Revision Process

The subset of 48 essays was submitted to W-Pal and the feedback was recorded. We then revised each essay based on the feedback suggested by W-Pal. To facilitate consistency and reduce the effects of personal style or bias, guidelines restricted the types of revisions the researchers could apply. Only feedback that was explicitly actionable within the essay was followed.

Specifically, feedback addressed one of four issues: finite elements, unbounded elements, spelling and grammar, or instruction. Finite elements were those such as thesis statements, topic sentences, or conclusions, which are usually present in specific amounts or locations (e.g., one topic sentence per body paragraph). Given feedback on a particular element, it was added if it was missing in the essay and edited if it was present but judged to be inadequate. However, if it was even minimally sufficient, it remained unedited.

Unbounded essay elements were those with no fixed quantity, such as examples, elaborations, or details. For this class, the requested element was simply added to the essay. One example or elaboration was added to each body paragraph, or one detail was added to each example.

W-Pal does not currently check spelling or grammar. Feedback only suggests that students double-check their own work. Revisions were limited to obvious sentence fragments, run-on sentences, capitalization errors, and misspelled words. However, misused words that were spelled correctly (e.g., substituting "their" for "there") were left unchanged.

Lastly, instructional suggestions were those that recommended students to review a lesson or complete a challenge for more information or practice. These recommendations were not implemented and no changes were made to the essay related to the recommended lesson.

Overall, essay revisions altered or added about one to four sentences per paragraph. For instance, when "more examples" were requested, a one-sentence example was added to each body paragraph. If another paragraph was requested, the new paragraph was three to four sentences. We attempted to mirror the writing style of the original draft, such as sentence length, word length, or particular example themes. The examples below show an original (A) and revised (B) body paragraph. The feedback encouraged the recipient to include a topic sentence and add elaboration. The specific changes made are in bold:

(A) The world in the present can find no true leaders. No one knows who to look up to in a time like this. Kids in America are not looking up to heroes anymore, but they are checking the gossip magazine everyday to see who hooked up with who. There is no real leadership in the political parties anymore because the congressmen are looking to become rich in any situation. They do not look at the leaders of the past for advice in a terrible situation that are country has now been put in. They are too busy becoming their own celebrities.

(B) The world in the present can find no true leaders. No one knows who to look up to in a time like this. Kids in America are not looking up to heroes anymore, but they are checking the gossip magazine everyday to see who hooked up with who. **Newspapers pay more attention to musicians' drug habits or love lives than to crises around the world.** There is no real leadership in the political parties**, Democrats or Republicans,** anymore because the congressmen are looking to become rich in any situation. They do not look at the leaders of the past for advice in a terrible **financial** situation that are country has now been put in. **Instead, they make selfish decisions based on what will make them popular or wealthier.** They are too busy becoming their own celebrities.

The methodology followed in modifying the essays has several limitations. First, it is confounded by the knowledge of the revisers. The revisers know the provenance of the essays and algorithm. Research bias could result in revisions not requested in the feedback. Second, although human ratings are available for the original drafts, such ratings are not available for the revisions. Thus, evaluations of essay improvement can only occur by comparing pre- and post-revision system scores. This is somewhat circular because the feedback is tied to such scores. However, the purpose of this usability testing method is to provide an efficient way to evaluate feedback for development purposes. The goal is to identify strengths and weaknesses of the feedback system early in the development process, prior to costly testing with students, rather than to establish accuracy or efficacy.

Results

Overall Accuracy of the Simplified Algorithm

Pearson correlations were computed between human scores and the scores assigned by the W-Pal algorithm for the entire original corpus (N = 201) prior to revision based on the W-Pal feedback. Pearson correlations yielded r = .48, p < .001. Perfect agreement was 20.9% and perfect + adjacent agreement was 65.2%. On average, the algorithm tended to underestimate scores by about one level. The mean human rating for essays was 3.4 (SD = 0.9) and the mean algorithm score was 2.5 (SD = 1.0). This difference was significant, t_{200} = -13.6, p < .001. Accuracy was not equal for all scoring levels. We calculated the mean absolute value discrepancy between scores for all levels (see Table 1), which revealed that the mean difference between algorithm and human scores was greater for higher quality essays, $F_{8,192}$ = 6.3, p < .001.

Human Score	*n*	Discrepancy	StDev
1.0	1	0.0	-
1.5	5	0.7	0.4
2.0	14	0.5	0.5
2.5	18	0.8	0.5
3.0	63	0.7	0.8
3.5	22	1.2	0.8
4.0	31	1.3	0.9
4.5	26	1.4	0.6
5.0	21	1.8	0.9

Table 1. Mean discrepancy (absolute value) between human scores and algorithm scores.

Given that the simplified algorithm relied on a small number of low-level text features, it is unsurprising that it was only marginally accurate at assigning essay scores. Furthermore, it makes sense that accuracy was poorer for higher quality essays. The writing properties that most likely discriminate among good essays (e.g., lexical sophistication and syntactic complexity; McNamara et al, 2010) were not included in the current simplified algorithm. While these more sophisticated indices will be implemented via Coh-Metrix in subsequent algorithms, the current evaluation remains useful because it provides an evaluation of the feedback mechanisms.

Essay Revision and Feedback

The original drafts of the 48 essays in the revision corpus were submitted to W-Pal. Fifteen essays fell below the length threshold and 9 essays fell below the paragraph structure threshold. No essays failed the legitimacy or relevance checks. Thus, although essays were sometimes too short or lacked structure, they were generally valid and on-topic. The mean score assigned by the system was 2.2, SD = 1.0. Table 2 reports the number of original essays at each holistic score level.

Following the revision procedures outlined above, each essay was revised and then resubmitted to W-Pal. No essay degraded in quality according to W-Pal, and none of the revised essays were below the length threshold. Three of the previously "too short" essays still lacked an adequate paragraph structure after revision. The mean W-Pal score for revised essays was 3.3, $SD = 0.6$. Thus, on average, revised essays received significantly higher scores than their original drafts, $t_{47} = 11.1$, $p < .001$. Table 2 reports the number of revised essays at each holistic score level.

	Revised Score				
Original Score	1.0	2.0	3.0	4.0	Total
1.0	0	4	11	0	15
2.0	0	0	11	1	12
3.0	0	0	5	12	17
4.0	0	0	0	4	4
Total	0	4	27	17	48

Table 2. Frequency and cross-tabulation of original and revised essays by holistic score.

Overall, the feedback supplied by the W-Pal system seemed usable in a manner that supported an increase in the scores assigned by W-Pal's algorithm. Our revision process was somewhat restrictive; we assumed that students would make rather minimal changes and ignore suggestions for additional work (e.g., creating an outline or watching a lesson), thought motivated students who utilize a broader scope of the feedback might improve further. Most importantly, this process yielded insights into issues to address as W-Pal feedback is refined in the next phase of development.

Observations

During the usability testing process, potential strengths and weaknesses of the feedback system were noted:

Feedback Quantity. W-Pal is intended to offer scaffolded feedback in which fundamental problems (e.g., paragraph structure) are addressed before higher-level problems (e.g., paragraph content). Feedback for "short" or "unstructured" essays was fairly brief. However, essays that surpassed these thresholds received more feedback on introduction, body, or conclusion paragraphs. Many essays triggered feedback on all three elements, which generated a daunting number of suggestions. This quantity of feedback seemed to undermine our scaffolding goal by targeting too many essay elements at once.

Repetitiveness. Over the course of several weeks, a semester, or school year, students will submit many drafts and revisions, often triggering the same feedback levels. Five versions of each feedback message were created to address the likely repetitiveness of system feedback. However, as we submitted essays for usability testing, it became clear that our "different" versions were not sufficiently distinct. For example, all suggestions for improving introductions tended to reference our "TAG" mnemonic: (T)hesis statement, (A)rgument preview, and (G)rab the reader's attention. Struggling writers who often make the same errors could grow frustrated with highly repetitive feedback messages.

Scope. Due to the simplified algorithm, our feedback necessarily addressed broad goals rather than specific errors. For example, we could not state that an essay "lacks a thesis." Instead, feedback described the importance or construction of a thesis, and posed questions or suggestions to prompt recipients to "double-check" their essay. Initially, the inability to give specific feedback seemed to be a liability. However, in practice, this approach may encourage a more holistic and metacognitive approach to essay revising. Instead of revising only to remove errors, writers might try to reconsider and refine the whole text.

Importantly, these issues potentially would have gone unnoticed if we had not, ourselves, gone through a systematic process of implementing the feedback from the perspective of a student. Thus, internal usability testing may be seen as a crucial step in the development process of AES systems.

Discussion

Automated essay scoring has become a popular tool that enables students to practice writing with individualized feedback. The automated scores generated by these systems are generally reliable. However, the efficacy of these systems – whether students improve in writing proficiency – may be strongly mediated by the nature and quality of the feedback they receive. We present a method to facilitate research on feedback within AES systems.

Authentic student essays were submitted to an AES for evaluation, and were then revised based on the actionable feedback provided. This internal usability evaluation indicated that feedback from an ITS improved the student essays after revision, albeit by the researchers. The scaffolded design of the feedback seemed to offset the limitations of the simplified algorithm. Most importantly, challenges and opportunities within the feedback design were identified, which will inform system development and future research. Despite limitations, the method offered insight into aspects of feedback accuracy, quantity, repetitiveness, and scope.

There remains a need for more research on the role of AES system feedback designed to guide students' writing development. This study prompts the following questions, which remain unanswered in the current literature.

What quantity of feedback is most beneficial in an AES learning environment? Should feedback be scaffolded or comprehensive? Research indicates that the amount of information delivered should be carefully managed to avoid overwhelming the learner (Kluger & DeNisi, 1996; Shute, 2008). However, most instructional AES systems provide comprehensive feedback on many topics for every essay draft.

What types of feedback are most beneficial in these environments? What modes or methods of feedback presentation are most effective and engaging? Students'

motivation to continue using the system might be influenced by the extent to which feedback messages are repetitive. Messages that adopt different approaches (e.g., examples, explanations, mnemonics, etc.) may help to stave off boredom. Moreover, varying combinations of feedback styles might better accommodate students who learn in different ways (Shute, 2008). Lastly, research suggests that effective feedback should support self-evaluation and self-regulation of learning (Butler & Winne, 1995). That is, feedback might be most effective when it encourages students to reflect on the writing process rather than on the specific list of revisions being implemented.

In conclusion, the current study was obviously only a first step in the usability process. First, we will revise the feedback system based on what we have learned. Then, in subsequent studies, we will examine the effectiveness of the feedback with student writers. We will examine which recommendations are implemented in revised drafts, and which are not, and whether the overall score improves. This method will provide authentic student data on feedback usability. Over time, combinations of internal and student-based usability tests of feedback design provide informative pictures of how to aid student learning via automated essay scoring.

Acknowledgments

This research was supported in part by the Institute for Education Sciences (IES R305A080589). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES.

References

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

Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment*, *4*(3). Retrieved Nov. 5, 2010 from www.jtla.org.

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

Butler, D. & Winne, P. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*, 245-281.

Craig, S., Driscoll, D., & Gholson, B. (2004). Constructing knowledge from dialog in an intelligent tutoring system: Interactive learning, vicarious learning, and pedagogical agents. *Journal of Educational Multimedia and Hypermedia*, *13*, 163-183.

Dikli, S. (2006). An overview of automated essay scoring of essays. *Journal of Technology, Learning, and Assessment*, *5*(1). Retrieved Nov. 5, 2010 from www.jtla.org.

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

Kellogg, R., Whiteford, A., & Quinlan, T. (2010). Does automated feedback help students learn to write? *Journal of Educational Computing Research*, *42*, 173-196.

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

Kluger, A. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254-284.

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., and Graesser, A. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*, 292-330.

McNamara, D., Raine, R., Roscoe, R., Crossley, S., Dai, J., Cai, Z., Renner, A., Brandon, R., Weston, J., Dempsey, K., Lam, D., Sullivan, S., Kim, L., Rus, V., Floyd, R., McCarthy, P., & Graesser, A. (in press). The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In P. McCarthy & C. Boonthum (eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution*. Hershey, PA: IGI Global.

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

Powers, D., Burstein, J., Chodorow, M., Fowles, M., & Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior*, *18*, 103-134.

Rock, J. (2007). *The impact of short-term use of Criterion on writing skills in 9th grade* (Research Report RR-07-07). Princeton, NJ: Educational Testing Service.

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.

Shermis, M. & Burstein, J., eds. (2003). *Automated essay scoring: A Cross-disciplinary perspective*. Mahwah, NJ: Erlbaum.

Shute, V. (2008). Focus on formative feedback. *Review of Educational Research*, *78*, 153-189.

Wang, J. & Brown, M. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment*, *6*(2). Retrieved Nov. 5, 2010 from www.jtla.org.

Warschauer, M. & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, *10*, 1-24.