Using Automatic Scoring Models to Detect Changes in Student Writing in an Intelligent Tutoring System

Scott A. Crossley[1], Rod Roscoe[3], and Danielle S. McNamara[2,3]

[1]Department of Applied Linguistics/ESL, Georgia State University
[2]Department of Psychology, [3]Learning Sciences Institute, Arizona State University

scrossley@gsu.edu, rod.roscoe@asu.edu, dsmcnamara1@gmail.com

Abstract

This study compares automated scoring increases and linguistic changes for student writers in two groups: a group that used an intelligent tutoring system embedded with an automated writing evaluation component (Writing Pal) and a group that used only the automated writing evaluation component. The primary goal is to examine automated scoring differences in both groups from pretest to posttest essays to investigate score gains and linguistic development. The study finds that both groups show significant increases in automated writing scores and significant development in lexical, syntactic, cohesion, and rhetorical features. However, the Writing-Pal group shows greater raw frequency gains (i.e., negative v. positive gains).

Introduction

A key measure of academic success is writing proficiency (Kellogg and Raulerson, 2007). However, attaining writing proficiency is often difficult and elusive (National Commission on Writing, NCW, 2003). One approach to improving writing skills is through the teaching of writing strategies, which facilitate task performance and accelerate skill acquisition. Common strategies used in writing instruction include planning, drafting, editing, summarizing, and revising. Teaching these strategies has proven effective in improving student writing, particularly for adolescent writers (Graham and Perin, 2007).

Strategy instruction is most successful when it is concrete, provides background knowledge for using the strategies, and provides opportunities to use the strategies through extended writing practice (i.e., opportunities to draft essays and revise them; Graham and Perin, 2007; Kellogg and Raulerson, 2007). In addition to opportunities to practice writing strategies, students need to receive feedback on their writing samples in order to improve their writing skills. There are generally two types of feedback: formative and summative. Formative feedback provides concrete guidance for student improvement (Shute, 2008) while summative feedback evaluates overall performance. Both types are important, but formative feedback has been identified as crucial for student development (McGarrell and Verbeem, 2007).

Automated Writing Evaluation

Automated writing evaluation (AWE) systems provide opportunities for students to practice writing and receive feedback in the classroom in the absence of a teacher. The feedback components in AWE systems are the major advantage such systems have over automated essay scoring (AES) systems (Grimes and Warschauer, 2010), which are only designed to provide accurate and reliable scores on essays or specific writing features such as grammar and mechanics. AES systems generally provide accurate scores to users that correlate with human judgments between .60 to .85 and report perfect agreement (i.e., exact match of human and computer scores) from 30-60% and adjacent agreement (i.e., within 1 point of the human score) from 85-99% (Attali and Burstein, 2006; McNamara, Crossley, and Roscoe, 2012; Rudner, Garcia, and Welch, 2006; Warschauer and Ware, 2006). However, accurate scoring on the part of AES systems does not appear to strongly relate to instructional efficacy with studies suggesting that students' essays improve in writing mechanics but not overall quality (Shermis, Burstein, and Bliss, 2004).

Like AES systems, AWE systems are also not without fault. While AWE systems can facilitate writing practice and improve motivation, users are skeptical about their scoring reliability (Grimes and Warschauer, 2010) and about the potential for AWE systems to overlook infrequent writing problems that, while rare, may be frequent to an individual writer. Lastly, AWE systems generally depend on summative feedback at the expense of formative feedback (Roscoe et al., 2012).

The Writing Pal

Intelligent tutoring systems (ITSs) adopt a pedagogical focus and are an alternative to strict AWE systems. The Writing Pal (W-Pal: McNamara et al., 2012) is one such

ITS, which provides writing strategy instruction to high school and entering college students. While most AWE systems focus on essay practice with some support instruction, W-Pal emphasizes strategy instruction and targeted strategy practice prior to whole-essay practice.

W-Pal offers writing strategies that cover three phases of the writing process: prewriting, drafting, and revising. Each of the writing phases is further subdivided into instructional modules. These modules include *Freewriting* and *Planning* (prewriting); *Introduction Building*, *Body Building*, and *Conclusion Building* (drafting); and *Paraphrasing*, *Cohesion Building*, and *Revising* (revising). An important component of W-Pal is that it incorporates a suite of games that target specific strategies for the writing processes above. The games allow students to practice the strategies in isolation before practicing the strategies in a complete essay. The games provide manageable sub-goals that defer the need to simultaneously coordinate the multiple tasks required during the writing process (e.g., Breetvelt, van den Bergh, and Rijlaarsdam, 1994). In W-Pal, students view lessons on each strategy, play practice games, and write practice essays for each of the modules.

Essay writing is an essential component of W-Pal. The system allows students to compose essays, and then provides holistic scores and automated, formative feedback based upon natural language input. The feedback in W-Pal depends on the W-Pal AWE system and focuses on strategies taught in the W-Pal lessons and practice games. For instance, if an essay is too short, the system provides feedback to the user about idea generation techniques such as freewriting. When an essay does not meet the paragraph threshold, the feedback system suggests techniques to plan and organize an essay more effectively including outlining and focusing on structural elements such as positions, arguments, and evidence. More specifically, students will be reminded to preview their thesis statements and arguments in the introduction paragraph, write concise topic sentences and present evidence in body paragraphs, and provide conclusion statements and restate the thesis in the concluding paragraph. Students are also given feedback on general revising, which includes condensing similar sentences, restructuring sentences, and improving cohesion. The feedback the student receives focuses on problem-identification and is stated in an impersonal and suggestive manner.

While W-Pal provides feedback and essay scoring, these elements are not the primary motivations for the system. Thus, unlike AWE systems, W-Pal was conceptualized as a system to provide instruction on writing strategies. While automated scoring is a key component of W-Pal, it is secondary to instruction (Roscoe et al., in press).

Automatically Scoring Essays in Writing Pal

The scoring system in W-Pal, like other AWE systems, assesses essay quality using a combination of computational linguistics and statistical modeling. However, unlike traditional scoring methods that rely on linear multiple regression models between text features and scores, the AWE in W-Pal uses hierarchical classification. Such hierarchical classification affords the opportunity to provide formative feedback at different conceptual levels on a variety of linguistic and rhetorical features.

The first step of the algorithm assumes that the largest difference between writers is between those who are able to produce enough information to have an acceptably structured essay. Hence, the first hierarchical categorization is a function of those who meet a threshold for number of words (250 words) and number of paragraphs (three paragraphs). In the following stages, the model assumes that essays that meet and do not meet the thresholds can be characterized by different linguistic features (as computed by Coh-Metrix and a variety of newer writing indices developed specifically for W-Pal). Such assumptions lead to a number of machine learning algorithms that are calculated separately for each group.

For instance essays in the group that do not meet basic thresholds are divided into high and low groups using linguistic features fed into a Discriminant Function Analysis (DFA). The essays in the low group are then further divided into low essays (scored a 1) and high essays (scored a 2). All remaining essays in the low group are classified as higher quality and assigned a score of 3. Essays that do meet the first threshold are also classified into low and high groups using linguistic features fed into a DFA. The low group is then further classified into low essays (scored a 2) and high essays (scored a 3). Essay in the high group are also divided into low essays (scored a 4) and high essays (scored a 5). Because a score of 6 is rare in the training corpora, a specific algorithm for a 6 was not developed. Instead, when the model determines that a given essay merits a "5," a secondary function examines the raw score computed by the algorithm. If the raw score for an essay is one standard deviation above the mean raw score, that essay is "upgraded" to a score of "6."

This approach was tested on a broad corpus of argumentative essays (N = 1243) written on 14 different prompts, by four different grade levels in four different timed conditions. All essays had been scored by at least two expert raters on a standardized scoring rubric with a range of 1 to 6. The derived scoring models provided exact accuracy of 55% and adjacent accuracy of 92%. The model was informed by 46 different linguistic, structural, rhetorical indices. The linguistic features relate to lexical sophistication, syntactic complexity, and cohesion. The structural features relate to text length, number of paragraphs, and comma use. The rhetorical features relate to semantic categories, conclusion statements, modal use, public and private verbs, and amplifiers.

The Current Study

In this study, we vary whether students receive more strategy instruction with half the writing practice (i.e., W-Pal) or receive double the writing practice with no instruction (i.e., an AWE system). We have two research questions in the current study: 1) Which mode of writing instruction leads to greater gains in automated scores (i.e., an ITS or AWE approach) and 2) Do the modes of writing instruction lead to differences in linguistic development? Our hypothesis is that both conditions will lead to increased essays scores and similar developments in linguistic competence even though students in the AWE condition write twice as many essays.

Method

We collected data from two groups of students. The first group interacted with the full W-Pal system described above. This group studied strategy lessons, completed brief quizzes, played practice games, and wrote and revised essays with feedback. The second group wrote and revised essays based on feedback from the W-Pal system, but did not interact with any other aspect of W-Pal. Both groups wrote pretest and posttest essays. In this study, we compare differences in the pretest and posttest essay scores as computed by the W-Pal scoring algorithm and differences in selected linguistic properties used in the W-Pal scoring algorithm.

Participants

For this study, we recruited 65 students from public high schools in the metro Phoenix area. Students ranged in age from 14 to 19 (M = 15.9, SD = 1.3) and ranged in grade level from 9 to 12 (M = 10.2, SD = 1.0). Of the 65 participants, 70.8% were female and 29.2% were male. Twenty-seven of the participants self-identified as English Language Learners (ELL). The remaining participants self-identified as native speakers of English (NS). Participants were divided into two conditions: the W-Pal condition (n = 33) or the Essay condition (n = 32). Of the 33 participants in the W-Pal condition, 23 self-identified as NSs and 10 self-identified as ELLs. Of the 32 participants in the Essay condition, complete data for 31 of the participants was available (post-test data from one student was not recorded due to a technical error). Of these 31 participants, 14 self-identified as NSs and 17 self-identified as ELLs. The data from these 64 participants are reported in this study.

Procedures

Students attended 10 sessions (1 session/day) over a 2-4 week period. Participants wrote a pretest essay during the first session and a posttest essay during the last session. The essays were written on two prompts (on the value of competition and on the role of image) counterbalanced across the pretest and posttest essays. In addition, the first and final sessions included assessments of reading comprehension, vocabulary, writing proficiency, strategy knowledge, and writing attitudes. However, these items were not analyzed as part of the current study.

Sessions 2-9 were devoted to training. The students in the W-Pal condition used the full W-Pal, including essay writing, instructional lessons, and practice games. W-Pal condition students wrote and revised one essay per session, and completed one instructional module. Essay condition students interacted only with the essay writing and automated feedback tools in W-Pal. These students wrote and revised two essays each session based on automated feedback. The time on task between conditions was equivalent overall.

Corpus and Scoring

The final corpus of essays used in this analysis comprised 128 essays written by the 64 participants for both pretest and posttest writing. The scoring algorithm discussed in the introduction to this paper was used to assign each pretest and posttest essay a holistic score between 1 and 6.

Selected Linguistic Properties

From the W-Pal scoring algorithm we selected a number of linguistic properties related to structure, lexical sophistication, syntactic complexity, cohesion, and rhetorical structure that were either the strongest predictors of a specific hierarchical categorization or significant predictors in two or more hierarchical categorizations. The structural features were number of words and number of paragraphs. The lexical features were lexical diversity (*D*; Malvern et al., 2004) and incidence of nominalizations. The syntactic feature was the incidence of Wh- relative clauses. The cohesion features were Latent Semantic Analysis (LSA; Landauer et al., 2007) paragraph to paragraph similarity and number of conjuncts (e.g., *however, thus*). The rhetorical structure was the incidence of conclusion statements.

Statistical Analysis

We first conducted *t*-tests between the essay scores assigned to NS and ELL participants during the pretest. This analysis allowed us to assess differences between the writing proficiency of the NS and ELL participants. We next conducted *t*-tests between the scores assigned to essays during the pretest between conditions (W-Pal and Essay conditions) to assess if participants differed in writing proficiency between the conditions. We then conducted *t*-tests between the calculated essays scores for the two prompts to ensure that prompt-based effects did not exist. We then conducted a repeated measures ANOVA between the algorithm scores and the selected linguistic features for the pretest and posttest essay scores. We included two between-subjects effects: condition (W-Pal or Essay) and writing proficiency level (as classified by their pretest scores). All participants that wrote an essay assigned a score of 1 or 2 in the pretest (*n* = 32) were labeled *low proficiency*. All participants that wrote an essay scored 3 or 4 (*n* = 32), were labeled *high proficiency*. These between-subjects effects were included to determine if differences in conditions and writing proficiency interacted with pretest and posttest scores. Lastly, we also conducted gain score comparisons between the W-Pal group and the Essay group using a *t*-test and chi-square analyses.

Pretest and Prompt Equivalence

Differences between NSs and ELL participants. There was no statistical difference in writing quality as measured by the scoring algorithm between ELL (M = 2.593, SD = .931) and NS participants (M = 2.351, SD = .887), (t = 1.051, DF = 62, p = .297). This finding indicates that the NS and ELL participants were of equal writing proficiency at the pretest.

Differences between conditions. There was no statistical difference in pretest writing quality for the participants in the W-Pal (M = 2.488, SD = 1.064) and the Essay condition (M = 2.419, SD = .721), (t = .286, DF = 62, p = .775). This finding indicates that the writers in both conditions were of equal writing proficiency at the pretest.

Prompt-based differences. There was no statistical difference between the writing prompts *Images* (M = 2.778, SD = .906) and *Competition* (M = 2.635, SD = 1.222) for all the essays in the corpus, (t = .894, DF = 62, p = .375). This finding indicates that there were no prompt-based writing effects for the assigned scores.

Pretest and Posttest Differences: Essay Scores

There was a significant main effect for essay score, F (1, 60) = 14.499, p < .001, hp^2 = .195 indicating that participants' essays were scored higher on the posttest than the pretest. There was not a significant interaction effect between test and condition, F (1, 60) = .003, p > .050, hp^2 = .000, indicating no differences in essay scores based on condition. There was, however, a significant interaction between test and writing proficiency level, F (1, 60) = 3.188, p < .050, hp^2 = .072, indicating that low proficiency writers showed greater gains than high proficiency writers. The three-way interaction was not significant, F (1, 60) = .639, p > .050, hp^2 = .011, indicating no differences in essay scores based on condition and proficiency level.

We conducted follow-up pairwise comparisons to evaluate differences between pretest and posttest scores for low and high proficiency writers within each condition. See Tables 1 and 2 for descriptive statistics for these analyses. Low proficiency writers in both the W-Pal condition (t = -3.503, df = 16, p < .010) and the Essay condition (t = -2.928, df = 15, p < .010) showed significant differences between their pretest scores and posttest essay scores. High proficiency writers in the W-Pal condition (t = -1.244, df = 16, p > .050) and the Essay condition (t = -.414, df = 14, p > .050) demonstrated no significant difference between their pretest and posttest essay scores.

Table 1
Essay scores for W-Pal condition: M (SD)

Essay	All participants	Low proficiency	High proficiency
Pretest	2.485 (1.064)	1.563 (0.512)	3.353 (0.606)
Posttest	3.030 (1.262)	2.313 (0.873)	3.706 (1.213)

Table 2
Essay scores for Essay condition: M (SD)

Essay	All participants	Low proficiency	High proficiency
Pretest	2.419 (0.720)	1.813 (0.403)	3.067 (0.258)
Posttest	3.000 (1.183)	2.813 (1.167)	3.200 (1.207)

Pre- and Posttest Differences: Linguistic Features

There was a significant main effect of test for all linguistic features except for D (lexical diversity; see Table 3) indicating that participants' produced a greater number of linguistic features related to essay quality when the posttest was compared to the pretest. No linguistic features showed a significant interaction between test and condition or a significant interaction between test and writing proficiency level. These results indicate no differences in linguistic features based on condition or proficiency level. The three-way interaction demonstrated an overall significant result for the LSA index, F (1, 60) = .639, p > .050, hp^2 = .011, and *the D* index, F (1, 60) = .639, p > .050, hp^2 = .011. Mean scores for these two indices (see Table 4) indicate that low proficiency writers in the W-Pal condition and high proficiency writers in the Essay condition showed greater gains in LSA scores. For lexical diversity D scores, low proficiency writers in the W-Pal condition and high proficiency writers in the Essay condition decreased while the scores increased for writers in the other conditions

Table 3
ANOVA results for linguistic features: Pretest and posttest

Index	*F*	*p*	*hp2*
Number of words	28.600	< .001	0.323
Number of paragraphs	32.818	< .001	0.354
LSA paragraph	18.919	< .001	0.240
D (lexical diversity)	0.148	> .050	0.002
Conjuncts	12.715	< .001	0.175
Nominalizations	4.561	< .050	0.071
Wh- relative clauses	6.857	< .050	0.103
Conclusion statements	15.869	< .001	0.209

Table 4
Three way interaction effects for LSA and *D* scores

Index	Condition	Proficiency	Time	Mean
LSA	W-Pal	Low	Pre	0.240
			Post	0.446
		High	Pre	0.388
			Post	0.448
	Essay	Low	Pre	0.306
			Post	0.381
		High	Pre	0.254
			Post	0.452
D	W-Pal	Low	Pre	81.980
			Post	74.127
		High	Pre	84.213
			Post	91.131
	Essay	Low	Pre	69.754
			Post	79.870
		High	Pre	87.964
			Post	83.451

Gain Differences

We conducted frequency gain analyses in which we coded each participant as showing negative, neutral, or positive gains between their pretest and posttest scores. We then conducted chi-square tests to assess gains or losses above expected frequencies for each condition. For the W-Pal condition, the chi-square test was significant, $X^2(2, n = 33) = 8.909$, $p < .05$, indicating that the observed number of participants that showed positive gains was greater than expected. For the Essay condition, the chi-square test was not significant, $X^2(2, n = 31) = 5.871$, $p > .05$, indicating that the observed number of participants that showed positive gains was not greater than expected. Table 5 contains the observed and expected frequencies.

Table 5
Observed and expected gain frequencies for conditions

Gain type	W-Pal observe	W-Pal expect	Essay observe	Essay expect
Negative	3	11	5	10.3
Neutral	16	11	16	10.3
Positive	14	11	10	10.3

Discussion and Conclusion

Whereas AWE researchers have primarily focused on score accuracy (Warschauer and Ware, 2006), there have been relatively few evaluations of student writing gains using AWE systems (e.g., Kellogg, Whiteford, and Quinlan, 2010) and few if any studies that have looked at automated score gains that result from using an ITS. This study takes a step in this direction by demonstrating that both the full W-Pal system and the W-Pal AWE system lead to increased gains in automated scores and accompanying linguistic features. In addition, the gain scores for both systems were stronger for lower proficiency writers than higher proficiency writers. In reference to frequency gain scores, this study found that a greater number of writers in the W-Pal condition showed unexpected gains as compared to the AWE system. Lastly, for all selected linguistic features except lexical diversity, students in both conditions showed gains in the expected direction indicating developing linguistic proficiency.

Overall, students who received writing instruction, played games, wrote essays, received feedback, and revised essays showed similar gains in automated scores as students that solely wrote essays, received feedback, and revised essays. The major difference between these two groups in terms of writing practice is that the students in W-Pal condition wrote and revised 8 essays over the course of the study, while the students in the Essay writing condition wrote and revised 16 essays. That the two groups show equal score gains indicates that instruction, game play, writing practice, and feedback may be as effective as writing practice mixed with feedback alone.

However, while the gains from both groups were equal in score differences, our frequency gains analysis showed that participants in the W-Pal condition showed more positive gains and fewer negative gains overall than the Essay condition. Forty-two percent of W-Pal participants showed positive gains from pretest to posttest scores and 9% showed negative gains. In contrast, 32% of participants showed positive gains and 16% showed negative gains in the Essay condition. We hypothesize that these differences may be related to task motivation and engagement. We assume that, as in past studies (e.g., Jackson and McNamara, 2011), students in the W-Pal condition were more engaged in the writing process and more motivated to write than students in the Essay condition. This engagement and motivation could lead to the positive gains we find in the W-Pal condition, but not the Essay condition; however, such assumptions require follow-up studies.

The findings from this study also indicate that explicit strategy instruction and game practice mixed with essay writing and feedback lead to similar linguistic gains as essay writing and feedback alone. In both conditions, writers produce longer essays that had a greater number of paragraphs, indicating that the structural elements of the essay improved. All students in this study also showed gains in the use of cohesion features. For instance, students used a greater number of conjuncts to explicitly mark logical relations between clauses and had greater semantic overlap between paragraphs indicating greater cohesion between structural elements. However, the three-way interaction for the LSA scores indicates that semantic similarity gains were greater for low proficiency writers in the W-Pal condition and high proficiency writers in the Essay condition. In reference to syntactic indices, all students showed an increased use of Wh relative clauses, demonstrating increased syntactic complexity and text descriptiveness. Rhetorically, students used a greater number of conclusion statements to explicitly indicate the final paragraph of the text. Findings for lexical sophistication features were mixed. Students did produce a greater number of nominalizations, indicating greater use of abstract terms. However, counter to expectation, the three-way interaction effect for lexical diversity indicated that low proficiency writers in the W-Pal and condition and high proficiency writers in the Essay condition produced essays with less lexical variation. When considered alongside the LSA findings, it may be that low proficiency writers in the W-Pal condition gain from explicit cohesion instruction. Such instruction could lead to the greater use of semantic co-referentiality seen and, conversely, to lower lexical variation (resulting from increased word overlap). High proficiency writers in the Essay condition may be advanced enough to make cohesive changes in their essays in the absence of instruction.

Thus, many of these changes likely occurred as a result of training, and as a result of feedback from the AWE algorithm, which focuses on improving essay structure, text cohesion, and rhetorical features. That similar improvements across conditions occurred is important considering that students in the Essay condition wrote twice as many essays as the W-Pal condition and thus received twice as much feedback from the AWE system. However, as with the essay scores, this finding likely indicates that explicit strategy training and game play lead to developments in linguistic knowledge and production that equals the additional feedback received in the essay condition. It is also interesting to note that these changes occurred in lexical and syntactic features although these features were not a focus of the feedback system. It is likely that these features serve other textual functions as well (i.e., relative clauses are linked to description and nominalizations are linked to summarization).

Overall, this study finds that students gain from both writing instruction that includes strategies, game play, practice, and feedback and more extensive writing practice that includes feedback alone. However, at a larger scale level, we see advantages for the W-Pal intervention. Follow-up studies should investigate differences in automated scores and linguistic changes and how these differences relate to task motivation and engagement.

Acknowledgments

This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES.

References

Attali, Y., and Burstein, J. 2006. Automated Essay Scoring with E-rater V.2. *Journal of Technology, Learning, and Assessment*, *4*3.

Breetvelt, I., van den Bergh, H, and Rijlaarsdam, G. 1994. Relations Between Writing Processes and Text Quality: When and How? *Cognition and Instruction, 12,* 103-123.

Graham, S. and Perin, D. 2007. A Meta-Analysis of Writing Instruction for Adolescent Students. *Journal of Educational Psychology*, *99*, 445-476.

Grimes, D. and Warschauer, M. 2010. Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation. *Journal of Technology, Learning, and Assessment*, *8*, 4-43.

Jackson, G.T., and McNamara, D.S. 2011. Motivational Impacts of a Game-Based Intelligent Tutoring System. In R. C. Murray and P. M. McCarthy Eds., *Proceedings of the 24th International Florida Artificial Intelligence Research Society FLAIRS Conference* pp. 519-524. Menlo Park, CA: AAAI Press.

Landauer, T., McNamara, D. S., Dennis, S., and Kintsch, W. Eds.. 2007. *Handbook of Latent Semantic Analysis.* Mahwah, NJ: Erlbaum.

Kellogg, R. and Raulerson, B. 2007. Improving the Writing Skills of College Students. *Psychonomic Bulletin and Review, 14*, 237-242.

Malvern, D. D. Richards, B. J., Chipere, N., and Duran, P. 2004. *Lexical Diversity and Language Development: Quantification and Assessment*. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

McGarrell, H. and Verbeem, J. 2007. Motivating Revision of Drafts through Formative Feedback. *ELT Journal, 61*, 228-236.

McNamara, D. S., Crossley, S. A., and Roscoe, R. in press. Natural Language Processing in an Intelligent Writing Strategy Tutoring System. *Behavior Research Methods*.

McNamara, D., Raine, R., Roscoe, R., Crossley, S., Jackson, G., Dai, J., Cai, Z., Renner, A., Brandon, R., Weston, J., Dempsey, K., Carney, D., Sullivan, S., Kim, L., Rus, V., Floyd, R., McCarthy, P., and Graesser, A. 2012. The Writing-Pal: Natural Language Algorithms to Support Intelligent Tutoring on Writing Strategies. In P. McCarthy and C. Boonthum-Denecke Eds., *Applied natural language processing and content analysis: Identification, investigation, and resolution* pp. 298-311. Hershey, P.A.: IGI Global.

Roscoe, R., Kugler, D., Crossley, S., Weston, J., and McNamara, D. S. 2012. Developing Pedagogically-Guided Threshold Algorithms for Intelligent Automated Essay Feedback. In P. McCarthy and G. Youngblood Eds., *Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference* pp. 466-471. Menlo Park, CA: The AAAI Press.

Rudner, L., Garcia, V., and Welch, C. 2006. An Evaluation of the IntelliMetric Essay Scoring System. *Journal of Technology, Learning, and Assessment*, 44.

Shute, V. 2008. Focus on Formative Feedback. *Review of Educational Research, 78*, 153-189.

Warschauer, M., and Ware, P. 2006. Automated Writing Evaluation: Defining the Classroom Research Agenda. *Language Teaching Research, 10*, 1-24.