Developing Pedagogically-Guided Threshold Algorithms for Intelligent Automated Essay Feedback

Rod D. Roscoe, Danica Kugler[1], Scott A. Crossley[3], Jennifer L. Weston[2], and Danielle S. McNamara[2]

[1] Dept. of English, [2]Dept. of Psychology, Learning Sciences Institute, Arizona State University, Tempe, AZ 85287
[3]Dept. of Applied Linguistics/ESL, Georgia State University, Atlanta, GA 30302
{rod.roscoe, danica.kugler, jennifer.weston, danielle.mcnamara}@asu.edu, sacrossley@gmail.com

Abstract

Current computer based tools for writing instruction show high scoring accuracy but uncertain instructional efficacy. One explanation is that these systems may not effectively communicate valid and appropriate formative feedback. In this paper, we describe an exploratory method for developing feedback algorithms that are grounded in writing pedagogy. The resulting threshold algorithms are shown to be meaningfully related to essay quality and informative regarding individualized, formative feedback for writers.

Introduction

A number of computer tools have been created to facilitate educators' ability to assess student writing and provide feedback. Grimes and Warschauer (2010) describe two kinds of systems: *automated essay scoring* (AES) and *automated writing evaluation* (AWE). AES systems strive to assign accurate and reliable scores to essays or specific writing features (e.g., mechanics). Scores are generated using various artificial intelligence (AI) methods, including statistical modeling, natural language processing (NLP), and Latent Semantic Analysis (LSA) (Shermis & Burstein, 2003). More recently, AES systems have been paired with educational and classroom management tools to create AWE systems that support scoring and writing instruction.

In this transition from automated scoring to writing instruction, issues of feedback design take on critical importance. Feedback is the primary means through which students can evaluate and improve their writing, and feedback quality has a profound impact. In particular, research has identified *individualized, formative feedback* as essential to students' writing proficiency development (McGarrell & Verbeem, 2007; Sommers, 1982). Formative feedback provides concrete guidance and methods for student *improvement* (Shute, 2008), such as strategies for generating and organizing ideas, arguments and evidence. In contrast, summative feedback evaluates *performance*, and may consist of grades and teacher critiques on spelling, lack of detail, weak arguments, and so on. Although both forms of feedback are useful, formative feedback is crucial for student growth because it renders the means and methods of such growth explicit. Thus, a critical question arises for developers of computer-based writing instruction tools: *How can we translate computational linguistic indices and measures into formative feedback that is valid and useful for developing writers?*

Whereas researchers have primarily focused on score accuracy (Warschauer & Ware, 2006), there have been relatively few evaluations of student improvement (e.g., Kellogg, Whiteford, & Quinlan, 2010) or the role of feedback (e.g., Roscoe, Varner, Cai, Weston, Crossley, & McNamara, 2011). Hence, in this paper, we explore and describe a method for developing pedagogically-guided algorithms that guide formative feedback in an intelligent tutor system (ITS) for writing.

Accuracy, Efficacy, and Validity of AES/AWE

Several commercial AWE systems are now available, such as *Criterion* (scored by *e rater*) from the Educational Testing Service, *MyAccess* (*IntelliMetric*) from Vantage Learning, *WriteToLearn* (*Intelligent Essay Assessor*) from Pearson Inc., and *WPP Online* (*PEG*) from Educational Record Bureau. Each system adopts a different scoring approach. For example, *Criterion* (Burstein, Chodorow, & Leacock, 2004) and *MyAccess* (Rudner, Garcia, & Welch, 2006) rely mainly on NLP and AI tools, whereas *Write To Learn* uses LSA (Landauer, Laham, & Foltz, 2003). These systems provide comprehensive feedback on essay traits such as spelling, grammar, mechanics, usage, and style.

Despite differing methods, scoring accuracy tends to be high; human and computer-based scores correlate around .80 to .85 (Landauer et al., 2003; Rudner et al., 2006, Warschauer & Ware, 2006). Several systems report "perfect" agreement (exact match of human and computer scores) from 40-60%, and "perfect+adjacent" agreement (human and computer scores within 1 point) from 90-100% (Attali & Burstein, 2006; Rudner et al., 2006).

Unfortunately, scoring accuracy does not appear to translate directly to instructional efficacy (Grimes & Warschauer, 2010). A handful of quantitative and qualitative evaluations have been reported, which suggest that students' essays improve in writing mechanics but not overall quality. For example, Shermis, Burstein, and Bliss (2004) compared state exam writing scores for over 1000 high school students, half of whom used *Criterion* and half of whom completed alternate classroom writing assignments. There was no difference in exam scores for the two groups, although *Criterion*-users wrote longer essays with fewer mechanics errors. Kellogg et al. (2010) manipulated how much feedback college students received from *Criterion* on three essays. Students received feedback on all three essays, one essay, or none of the essays. There was no condition effect on essay scores; although students who received more feedback displayed fewer grammar, mechanics, usage, and style errors in their revisions.

Several qualitative evaluations of *MyAccess* have been published (Grimes & Warschauer, 2010). Students and teachers reported that *MyAccess* facilitated writing practice and improved motivation, but users were skeptical about scoring reliability. Many teachers (39%) disagreed with the statement, "MyAccess gives fair and accurate scores." Students also reported feeling overwhelmed by the quantity of feedback provided. Teachers described the need to generate supplementary materials to help students navigate the "pages of suggestions" given by the system.

Overall, there may be a disconnect between scoring accuracy and instructional efficacy. One concern may be the validity of the automated scores (e.g., Clauser, Kane, & Swanson, 2002). Purely statistical methods for algorithm-generation may unintentionally overlook writing problems that occur infrequently. Feedback that is based only on common predictors, or predictors that cut across large aggregates, may neglect serious problems displayed by *individual students*. We must be able to provide formative feedback to students even for uncommon writing problems.

Formative Feedback in an ITS for Writing

Our research group is developing the Writing Pal (W-Pal), an intelligent tutoring system that offers strategy instruction, practice, and feedback for developing writers (McNamara et al., 2011). Currently, the system focuses on the genre of persuasive writing. Strategy instruction is delivered via Strategy Modules that address multiple phases of writing: prewriting (Freewriting and Planning), drafting (Introduction Building, Body Building, and Conclusion Building), and revising (Paraphrasing, Cohesion Building, and overall Revising). Each module contains instructional videos narrated by a pedagogical agent, and mini-games that enable game-based practice. Students also practice writing essays, which receive holistic scores and formative feedback generated by a series of algorithms.

W-Pal's approach differs from alternative systems, such as *Computer Tutor for Writing* (Rowley & Meyer, 2003) or *Glosser* (Villalon, Kearney, Calvo, & Reimann, 2008), which offer strategy help or reflective questions embedded within the writing interface. These systems scaffold students *as they write* to produce texts that are relevant, organized, and meet key writing goals. In contrast, W-Pal provides in-depth strategy training and practice *prior to writing* paired with formative, *automated strategy feedback* on students' authored essays or excerpts. W-Pal does not interrupt students during the writing process.

Formative feedback design is a key aspect of W-Pal development because of our focus on strategy instruction. Our algorithms must be sensitive to both overall essay quality and individual students' writing strategies. Such demands have revealed the inadequacy of feedback engines that are too tightly coupled with scoring engines. Although we can achieve reasonable percent agreement between human and computer scores, it can be challenging to translate scoring algorithm outputs into pedagogically-valid feedback. Statistical models do not always capture the unique set of problems displayed in an essay, and scoring indices may not map on to the kinds of a feedback a writing instructor would offer. Moreover, it is not always clear when to give feedback. Receiving feedback on every essay feature is overwhelming (Grimes & Warschauer, 2010). Alternatively, we try to target a subset of major problems for each student, but how can we determine thresholds to govern such responses?

We are currently exploring alternative methods for developing feedback algorithms. These algorithms are not derived from scoring algorithms. Instead, computational linguistic measures are mapped onto guidelines from writing style resources (e.g., Hacker, 2009). We first identify pedagogical principles, such as "Use specific examples," and then connect these writing features to specific computational indices that capture the desired construct(s). Subsequent analyses produce threshold values for the target variables, which govern when and how to respond to problems in students' essays. Ultimately, we expect that the resulting algorithms will have high intrinsic pedagogical validity, which should contribute to greater effectiveness for improving students' writing proficiency.

Algorithm Development

Essay Corpus

From prior research on writing and essay analysis, we obtained a corpus of 526 essays written by high school or freshman college students. All essays were written based on SAT-style persuasive essay prompts with a 25-minute time limit. Prompts addressed a variety of topics: originality (n = 177), heroes (n = 138), choices (n = 70), optimism (n = 56), memories (n = 45), and change (n = 40). All essays had been scored by trained, reliable human raters using a SAT-based rubric (i.e., a 6-point scale).

Selection of Linguistic Indices

Writing style guides provide numerous suggestions for persuasive essay writing. Writers are instructed to "support your central claim and any subordinate claims with evidence: facts, statistics, examples, illustrations, expert opinion, and so on" (Hacker, 2009, p. 363), and taught to use specific evidence, such as references to particular dates and people. Guides also advise authors regarding word choice, such as using "exact words" and "specific, concrete nouns [to] express meaning more vividly" (Hacker, 2009, pp. 138-140). Similarly, writers may be told to maintain objectivity by writing in the 3rd person perspective.

The scoring rubrics used to assess writing are another source of information about writing pedagogy (de la Paz, 2009). The SAT essay scoring rubric has six levels corresponding to a holistic 1-to-6 rating. High-scoring essays use "clearly appropriate examples, reasons, and other evidence" and exhibit "skillful use of language, using a varied, accurate, and apt vocabulary." In contrast, low-scoring essays provide "little or no evidence" and display "fundamental errors in vocabulary" (College Board, 2011).

Example Writing Guideline	Linguistic Measure
Vocabulary	
"Use larger words"	Mean syllables per word
"Use varied words"	Lexical diversity
"Use concrete words"	Mean word concreteness
"Use specific words"	Mean word hypernymy
Evidence	
"Offer ample evidence"	Total number of words
"Provide examples"	Exemplification n grams
"Use specific examples"	Date time references
"Avoid uncertainty"	Possibility words
Perspective	
"Maintain objectivity"	1st, 2nd, 3rd person pronouns
Transitions	
"Clearly connect ideas"	Additive, causal, logical, and negation connectives

Table 1. Writing features and associated linguistic measures analyzed to develop feedback thresholds.

From these sources, we compiled common writing guidelines (Table 1). Next, these guidelines were mapped onto linguistic indices that might capture the construct. We used Coh-Metrix (Graesser & McNamara, 2011; McNamara & Graesser, 2011), an automated tool that comprises many indices, including basic features (e.g., word count), lexical sophistication (e.g., concreteness), syntactic complexity (e.g., syntactic similarity), referential cohesion (e.g., argument overlap), and deep cohesion (e.g., causal verb overlap). Additional tools assessed the occurrence of automatically-determined or pre-determined or key words, such as n-grams related to rhetorical functions (Crossley, Roscoe, & McNamara, 2011). A small subset of guidelines and 15 variables were selected to facilitate exploration of this algorithm-generation process.

Vocabulary. Word size was assessed via the mean number of syllables per word in the essay. Word variation was assessed by the occurrence of unique words relative to essay length. Word concreteness assessed the extent which words in the essay referred to concrete sensory experiences (e.g., *apple*) versus abstract concepts (e.g., *truth*). Word hypernymy assessed the mean level of specificity of words in the text. For example, *chair* is more specific than *furniture*, but *recliner* is more specific than *chair*.

Evidence. Word count was used as a rough measure of elaboration; short essays are necessarily underdeveloped. Use of examples was assessed by the incidence of exemplification n-grams, such as *for example*, which often mark the beginning of examples. The incidence of date-time references (e.g., *November*) measured use of specific evidence and examples. Incidence of possibility words (e.g., *might*) measured uncertainty in an essay.

Perspective. The incidence of first, second, and third-person pronouns were used to assess objectivity.

Transitions. The incidence of additive (e.g., *moreover*), causal (e.g., *as a result*), logical (e.g., *therefore*), and negation (e.g., *on the other hand*) connective phrases assessed how students linked ideas with transitions.

Establishing Feedback Thresholds

Writing is an ill-defined domain, and it can be difficult to determine precisely whether a problem has occurred and merits feedback. One method is to establish thresholds for selected linguistic variables such that only essays that fall above or below the threshold receive feedback.

To establish objective, quantitative thresholds, we employed a *binning process* that grouped essays into four categories based on mean and standard deviation values for each variable. Essays that were *within one standard deviation above* the mean were placed in the "High" bin for the variable, and essays that were *more than one standard deviation above* the mean were placed in the "Highest" bin. Similarly, essays that were *within one standard deviation below* the mean were grouped in the

"Low" bin, and essays that were *less than one standard deviation below* the mean were placed in the "Lowest" bin.

Algorithm Validation

Initial validation of the algorithms occurred in two stages. First, for each variable, we assessed whether essay scores differed based on bin categorization. Second, we read example essays from low-scoring bins to assess whether essays exhibited the desired target problem.

Essay Scores Based on Bin Categorization

Table 2 summarizes mean essay scores by bin category. One-way ANOVAs revealed a significant main effect of bin categorization for most variables. In some cases, only the Lowest or Highest bin scored significantly different than the other bins. In other cases, the pattern appeared to be curvilinear: both the Highest and Lowest bins scored poorly compared to the High and Low bins.

	Bin Category				
Linguistic Measure	1	2	3	4	*F* value
Vocabulary					
Avg. syllables/word	**2.3**	2.9	3.2	3.3	16.99[a]
Lexical diversity	**1.9**	2.9	3.1	3.4	20.46[a]
Word concreteness	2.9	3.1	3.0	**2.5**	4.94[b]
Word hypernymy	**2.1**	2.9	3.2	3.4	24.61[a]
Evidence					
Number of words	**1.6**	2.7	3.4	3.8	116.71[a]
Exemplification	**2.4**	3.3	3.2	**2.7**	22.40[a]
Date time		**2.8**	3.2	3.2	6.17[b]
Possibility words	3.0	3.1	2.8	**2.7**	3.78[c]
Voice					
1st person pronouns	3.0	3.0	2.7	2.9	2.18
2nd person pronouns		3.1	2.6	**2.0**	25.25[a]
3rd person pronouns	**2.5**	3.0	3.0	3.1	5.61[a]
Transitions					
Additive connectives	**2.7**	3.0	3.1	2.9	2.24
Causal connectives	2.9	3.0	3.0	**2.6**	2.10
Logical connectives	2.8	3.1	3.0	**2.4**	6.61[a]
Negation connectives	2.9	3.0	3.0	**2.6**	3.27[b]

Note. 1 Lowest, 2 Low, 3 High, 4 Highest.
[a]$p < .001$. [b]$p < .01$. [c]$p < .05$.

Table 2. Mean essay scores based on bin categorizations.

Vocabulary. Essays that contained shorter words, less variation, and less specificity received lower scores. This pattern is consistent with instruction that teaches writers to use sophisticated and precise vocabulary. In contrast, essays with higher word concreteness received lower scores. Although guides suggest use of concrete wording to express meaning more clearly, perhaps essays that are too concrete are judged as lacking deep or critical thinking.

Evidence. Essays that were very short, lacked specific references to dates and times, or overused hedging words received lower scores. As expected, writers who relied on vague and uncertain evidence were judged as less proficient. Interestingly, the pattern for exemplification n-grams was curvilinear. Writers who rarely used exemplification markers received lower scores, perhaps because their essays lacked examples. However, overuse of these markers also negatively impacted scores. This may have occurred when students provided too many examples but did not develop examples sufficiently.

Perspective. No significant differences were observed for first-person pronoun use. This measure may not have captured the ways that first-person perspective was used well or poorly in an essay. Writers who can effectively use personal experiences to make a point will likely earn better scores, whereas writers who rely on unsupported opinion statements (e.g., *I think*) may earn lower scores. Results for second- and third-person pronouns were clearer. Essays written in the more objective third person received higher scores, whereas overuse of the informal second-person (*you/your*) perspective resulted in very low scores.

Transitions. No significant differences were observed for additive connectives or causal connectives, although the data suggest that overuse of causal connectives could negatively impact scores. Frequent use of logical or negation connectives was associated with lower scores. When logical connectives (e.g., *therefore*) are overused, this may be indicative of an essay where many claims are made but not supported by evidence. Finally, frequent negations may indicate essays where writers contradict themselves or seem unable to adopt a clear position.

Analysis of Sample Essays

Although essays categorized based on selected variables differed in score, it cannot be directly assumed that the variables captured the target constructs. Examining example essays from low-scoring bins helps to determine whether a meaningful writing problem was detected, and what feedback might be offered. The following sections provide examples from low-scoring *hypernymy*, *second person pronouns*, *exemplification*, and *logical connectives bins*. Spelling and punctuation were not corrected in these excerpts. Italics were added to highlight key phrases.

Hypernymy. Mean word hypernymy was intended to detect essays that used vague wording. The excerpt below was extracted from an essay in the Lowest bin. As expected, the writer used many vague words and failed to specify many concepts. For example, the first sentence did not clarify the comparison being made, and the second sentence never elaborated on the kinds of "problems" that might be seen or helped.

Formative feedback for this writer might include strategies for recognizing and clarifying undefined

referents, and could include instruction on how to elaborate vague sentences with explanations and examples.

> Being realistic is *more effective*. when people are realistic they can *see and help problems*. Insted of ignoreing them. In the real world *not everything* is perfect and people can't go out *beliveing that.*

Second person Pronouns. The incidence of second-person pronouns was intended to assess the objectivity employed by the writer. Frequent use of the second-person is associated with informal communication and is often not appropriate for academic writing. In the excerpt below, the tone is highly informal. Feedback for this author could instruct the writer about how to use other perspectives to project a more objective stance.

Another trend noted for many essays in the Highest second-person bin was the frequency of speculative if-then claims, such as "if you don't work for it you won't learn." Rather than building a case to support their claims, writers merely made a series of pronouncements directed at the reader. Further exploration of this pattern may show that writers who abuse second-person pronouns may also benefit from formative feedback regarding how and why to avoid speculative arguments.

> *You* will gain confidence, and *you* can get what *you* want. *You* learn from what *you* do but If *you* don't work for it *you* won't learn anything. It's still important to have optimism though. if *you* don't look forward to the future or look for the good things *your* life will be terrible.

Exemplification n grams. Exemplification was intended to measure whether writers incorporated examples. Our analyses found that both the Lowest and Highest bins for this variable received lower essay scores. The first excerpt below was extracted from the Lowest bin essays. No exemplification n-grams were observed; in fact, the essay did not contain nor develop any specific examples. Such writers may benefit from formative feedback on strategies for generating and elaborating examples.

> The human kind is a very jealous and ambicious species It is very nice to see by how much we have changed our ways of living, but there are some people that are unable to change their way of living, so they try to do whatever they can to change it even if it means hurting or even kill other people.

The second excerpt was extracted from the Highest exemplification bin. This essay contained examples, but did not elaborate examples with pertinent details. These writers likely do not need feedback on how to brainstorm examples. However, these students may need to learn strategies for expanding upon their examples and clarifying how the examples contribute to the argument. For example, the excerpt below might be improved by offering more specific details (e.g., names) about artists who have borrowed and built upon each others' work.

> Artist *come to mind* when creativity is mentioned. Creating something that no one else has done before can be hard, but using some ideas are not a bad thing. *For example*, if a artist use some ideas of *another* artist and then create something it still makes that *particular* piece of work original in some way.

Logical connectives. Logical connectives can signal the flow of the writer's arguments. Essays in the Highest bin for this variable received lower scores. In the excerpt, the author used logical connectives somewhat inappropriately. The writer presented a story about teenagers watching television and copying others. No logical conclusions can be deduced because no evidence was given. The incidence of logical connectives was inflated, however, by the use of the term *or*. The writer overuses the *he or she* construction to avoid gender-biased language, which results in awkward sentences. This pattern may have negatively impacted the score more than the misuse of phrases such as *therefore* and *actually*. In this case, it is unclear what formative feedback might be offered based solely on bin categorization for the logical connectives variable.

> Today, many teenagers watch television and see various styles and trends from famous people such as rappers and different celebrities. *Therefore*, he *or* she attempts to go out and be original and buy different things to say that he *or* she already came up with a certain style, but *actually*, the style that he *or* she is attempting has already been thought of *or* has been previously originated.

Conclusion

The delivery of valid formative feedback is a critical aim for any AWE intended to improve students' writing. Such improvement depends upon how well writing expectations and strategies are communicated to learners. Inappropriate feedback that is invalid, overwhelming, or too vague may explain why some AWE systems currently show high scoring accuracy but dubious instructional efficacy.

In this paper, we described a method for developing pedagogically-guided feedback algorithms. Rather than basing feedback algorithms on existing scoring algorithms, feedback thresholds were created by mapping guidelines from composition instruction materials to quantifiable linguistic indices. Many of our threshold-based essay categorizations were shown to capture meaningful distinctions in essay quality, and to provide insight into the potential problems that individual writers may experience.

These thresholds can be used to guide feedback that targets essays with specific problems. For instance, very short essays likely contain poorly developed ideas. Essays that fall into the Lowest word count bin (i.e., below 191 words) could receive formative feedback about idea generation strategies. This threshold process allows us to

confidently target individual essays for specific writing feedback – the student's work differs from other essays by at least a full standard deviation on the threshold variable.

The procedures described here were exploratory. Future work may be improved in several ways. First, threshold values may be more robust if obtained from a larger essay corpus, thus reducing the likelihood of overfitting. Second, threshold values may differ for high school students versus college students. Writers at different levels of development exhibit different linguistic and textual patterns. To create formative feedback that is sensitive to individual writers, the age of the writer may need to be taken into account. Similar variance may also be observed across different types of writing assignments and genres. Some prompts may favor different methods of effective argumentation.

Our review of essays within low-scoring bin categories also revealed nuances that were not apparent from patterns of essay scores. For example, essays with low hypernymy not only used vague words, but suffered from vague examples and arguments throughout. Similarly, writers who relied heavily on the second-person perspective often provided only speculative, hypothetical arguments that were not supported by evidence. In other words, essays displayed multiple problems at the same time. Further algorithm development could further address the "co-morbidity" of writing problems by developing thresholds based on *combinations* of linguistic measures. For example, assessing both second-person pronoun usage and logical connectives together might result in a reliable detector of "speculative reasoning," which could inform formative feedback about argumentation strategies.

In sum, the methods presented here have high potential for quickly and efficiently generating flexible algorithms for feedback in computer-based writing instruction. Our expectation is that this will improve the effectiveness of such tools for students' writing development.

Acknowledgments

This research was supported in part by the Institute for Education Sciences (IES R305A080589).

References

Attali, Y. & Burstein, J. (2006). Automated essay scoring with e rater V.2. *Journal of Technology, Learning, and Assessment*, *4,* 3 30.

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

Clauser, B., Kane, M. & Swanson, D. (2002). Validity issues for performance based tests scored with computer automated scoring systems. *Applied Measurement in Education*, *15*, 413 432.

College Board (2011). Essay scoring guide: A framework for scoring SAT essays. Retrieved November 15, 2011, from http://professionals.collegeboard.com/testing/sat reasoning/scores/essay/guide

Crossley, A., Roscoe, R., & McNamara, D. (2011). Predicting human scores of essay quality using computational indices of linguistic and textual features. *Proc. of the 15th Intl. Conference on Artificial Intelligence in Education.* Auckland, NZ: AIED.

de la Paz, S. (2009). Rubrics: Heuristics for developing writing strategies. *Assessment for Effective Intervention*, *34*, 134 146.

Graesser, A. & McNamara, D. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 2*, 371 398.

Grimes, D. & 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.

Hacker, D. (2009). *Rules for Writers: 6th Edition*. Boston, MA: Bedford/St. Martin's.

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

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

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

McGarrell, H., & Verbeem, J. (2007). Motivating revision of drafts through formative feedback. *ELT Journal. 61*, 228 236.

McNamara, D. & Graesser, A. (2011). Coh Metrix: An automated tool for theoretical and applied natural language processing. In P. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution*. Hershey, PA: IGI Global.

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. (2011). 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.

Roscoe, R., Varner, L., Cai, Z., Weston, J.., Crossley, S., & McNamara, D. (2011). Internal usability testing of automated essay feedback in an intelligent writing tutor. In R. C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 543 548). Menlo Park, CA: AAAI Press.

Rudner, L., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric essay scoring system. *Journal of Technology, Learning, and Assessment*, *4,* 3 21.

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

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

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

Sommers, N. (1982). Responding to student writing. *College Composition and Communication, 33,* 148 156.

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