Developing pedagogically-guided algorithms for intelligent writing feedback

Rod D. Roscoe*

Human and Environment Systems Department,
Learning Sciences Institute,
Arizona State University,
7271 E. Sonoran Arroyo Mall, 150D Santa Catalina,
Mesa, AZ 85212, USA
E-mail: rod.roscoe@asu.edu
*Corresponding author

Laura K. Varner

Department of Psychology,
Learning Sciences Institute,
Arizona State University,
P.O. Box 872111, Tempe, AZ 85287, USA
E-mail: laura.varner@asu.edu

Scott A. Crossley

Department of Applied Linguistics/ESL,
Georgia State University,
34 Peachtree St. Suite 1200, One Park Tower Building,
Atlanta, GA 30303, USA
E-mail: sacrossley@gmail.com

Danielle S. McNamara

Department of Psychology,
Learning Sciences Institute,
Arizona State University,
P.O. Box 872111, Tempe, AZ 85287, USA
E-mail: danielle.mcnamara@asu.edu

Abstract: Various computer tools have been developed to support educators' assessment of student writing, including automated essay scoring and automated writing evaluation systems. Research demonstrates that these systems exhibit relatively high scoring accuracy but uncertain instructional efficacy. Students' writing proficiency does not necessarily improve as a result of interacting with the software. One question is whether these systems offer appropriate or sufficient formative feedback to students about their writing. To motivate further research in this area, we present a straightforward methodology for constructing automated feedback algorithms that are grounded in writing pedagogy and assessment. The resulting threshold algorithms are demonstrated to be meaningfully related to essay quality and informative regarding individualised, formative feedback for writers. Potential applications and extensions of this methodology are discussed.

Keywords: automated essay scoring; AES; automated writing evaluation; AWE; intelligent tutoring systems; formative feedback; natural language processing; NLP; learning technologies; writing pedagogy.

Reference to this paper should be made as follows: Roscoe, R.D., Varner, L.K., Crossley, S.A. and McNamara, D.S. (2013) 'Developing pedagogically-guided algorithms for intelligent writing feedback', *Int. J. Learning Technology*, Vol. 8, No. 4, pp.362–381.

Biographical notes: Rod D. Roscoe is an Assistant Professor in the Department of Human and Environmental Systems, Cognitive Science and Engineering Programme, and the Learning Sciences Institute at Arizona State University. His research examines self-regulated learning processes in formal and informal contexts, and examines how these processes can be facilitated via adaptive technology, instruction, and peer support.

Laura K. Varner is a PhD student in Psychology and the Learning Sciences Institute at Arizona State University. Her research examines the cognitive processes and abilities underlying reading and writing proficiency, and also considers the impact of these factors in second language learning.

Scott A. Crossley is an Assistant Professor of Applied Linguistics and ESL at Georgia State University. His research focuses on corpus linguistics and the application of computational tools in second language learning and text comprehensibility.

Danielle S. McNamara is a Senior Research Scientist at the Learning Sciences Institute and a Professor of Psychology at Arizona State University. The overarching theme of her research is to better understand cognitive processes involved in comprehension, writing, knowledge acquisition, motivation, and memory, and to apply that understanding to educational practice by developing and testing educational technologies.

This paper is a revised and expanded version of a paper entitled 'Developing pedagogically-guided threshold algorithms for intelligent automated essay feedback' presented at the 25th International Florida Artificial Intelligence Research Society (FLAIRS) Conference, Marco Island, FL, USA, May 2012.

1 Introduction

As computer-based writing instruction gains prominence, issues of feedback design take on crucial importance. Feedback is a fundamental means through which students can evaluate and improve their writing. Automated feedback to students on writing might be conveyed in at least two different ways. First, students might receive summative feedback or an overall score, such as *poor* (e.g., with a corresponding score of 1) to *excellent* (e.g., with a corresponding score of 6). Second, students might receive formative feedback on the quality of the essay. These types of feedback manifest in two different types of technologies designed for writing assessment and instruction: *automated essay* *scoring* (AES) and *automated writing evaluation* (AWE). The purpose of AES is to assign accurate grades to essays, primarily to facilitate the scoring of large numbers of essays from standardised tests. Such scores are determined through artificial intelligence (AI) methods such as statistical modelling, natural language processing (NLP), and latent semantic analysis (LSA) (Dikli, 2006; Graesser and McNamara, 2012; McNamara et al., 2013; Shermis and Burstein, 2003). More recently, AES has been integrated with instructional materials and classroom management tools to create AWE systems that support both scoring and writing instruction. A variety of AWE systems are now available, including *Criterion* (scored by the *e-rater* AES system) from the educational testing service, *MyAccess* (scored by *IntelliMetric*) from vantage learning, *WriteToLearn* (scored by *Intelligent Essay Assessor*) from Pearson Inc., and *WPP Online* (scored by *PEG*) from Educational Record Bureau.

While AES is vital and has multiple benefits, *individualised, formative feedback* is essential to students' writing proficiency development (McGarrell and Verbeem, 2007; Sommers, 1982). Formative feedback offers learners concrete guidance and methods for improvement (Shute, 2008), such as strategies for evaluating and presenting objective evidence. In contrast, summative feedback evaluates performance and may comprise teacher grades and critiques on grammar or other errors. Although both forms of feedback are beneficial, formative feedback is important for student growth because it clearly communicates the steps necessary for improvement. Consequently, a core question for computer-based writing instruction tools is *how to translate computational linguistic data into formative feedback that is valid and useful for developing writers.*

Although studies of AES and AWE have demonstrated accurate scoring (e.g., Dikli, 2006; Warschauer and Ware, 2006), relatively few studies have assessed how student writing improves due to interacting with AWE systems (e.g., Kellogg et al., 2010) or the influence of feedback (e.g., Roscoe et al., 2011). In this study, we consider a novel method for developing pedagogically-guided algorithms with the potential to guide formative feedback in an intelligent tutoring system for writing. This method explicitly links feedback algorithms to principles of writing communicated in writing standards, scoring rubrics, and writing style guides. In sharing this work, a secondary goal is to stimulate further research on the design and evaluation of automated formative feedback.

2 Effects of automated scoring and writing evaluation

Each AWE system adopts a somewhat different scoring method. For example, *Criterion* (Burstein et al., 2004) and *MyAccess* (Rudner et al., 2006) have used NLP and AI methods. A corpus of essays is first scored by human raters and then automatically analysed along multiple linguistic features. Statistical analyses extract features that discriminate between higher and lower-quality essays and detect essay errors. Finally, statistical models combine the extracted linguistic properties into algorithms that assign grades. In contrast, *WriteToLearn* (Landauer et al., 2003) uses LSA. One assumption of LSA is that word meanings are partly established by their cooccurrence with other words, and thus texts can be semantically represented by a multidimensional word-by-context matrix. Singular value decomposition reduces the number of dimensions to capture semantic structure. To assign grades, student essays are scored based on their semantic similarity to a benchmark corpus of pre-scored essays.

Correlations between human and computer-assigned scores – typically using a SAT-like rubric on a six-point scale – correlate around .80–.85 (Landauer et al., 2003; Rudner et al., 2006; Warschauer and Ware, 2006). Several studies report 'perfect agreement' (i.e., exact match between human and automated scores) from 40%–60% and 'adjacent agreement' (i.e., human and computer scores are within one point) from 90%–100% (Attali and Burstein, 2006; Dikli, 2006; Rudner et al., 2006). However, accurate scoring does not necessarily guarantee that students' writing will improve (Grimes and Warschauer, 2010). Available research suggests that students' gains in writing are limited to mechanical details (e.g., spelling and punctuation) rather than overall writing quality.

For example, Shermis et al. (2004) compared state exam writing scores for over 1000 high school students, half of whom used *Criterion* and half of whom completed typical writing assignments. *Criterion* provided feedback tied to its scoring algorithms, which produced ratings related to errors in grammar, usage, mechanics, organisation, and style. The two groups did not differ in scores although *Criterion*-users wrote longer essays with fewer mechanical errors. Similarly, Attali (2004) analysed over 9,000 essays written by *Criterion* users (6th–12th grade) to compare original and final drafts. Final drafts earned higher automated scores (effect size d = .47), were longer, and contained fewer mechanical errors. Essays also seemed to improve in stating and supporting main ideas, and incorporating conclusions. Attali (2004) attributed these gains to receiving *Criterion* feedback. Importantly, however, a non-*Criterion* comparison condition was not included, and thus maturation and teacher instruction may have played crucial roles in students' gains. Finally, Kellogg et al. (2010) manipulated the amount of feedback undergraduates received from *Criterion* on three essays. Students received feedback on all essays (*continuous feedback*), one essay (*intermittent feedback*), or none of their essays (*no feedback*). Results indicated that revised essays received higher *Criterion* scores than original drafts; there was no effect of condition. Essay quality improved regardless of whether automated feedback was provided, although continuous feedback was more effective than no feedback in reducing grammar, mechanics, and style errors.

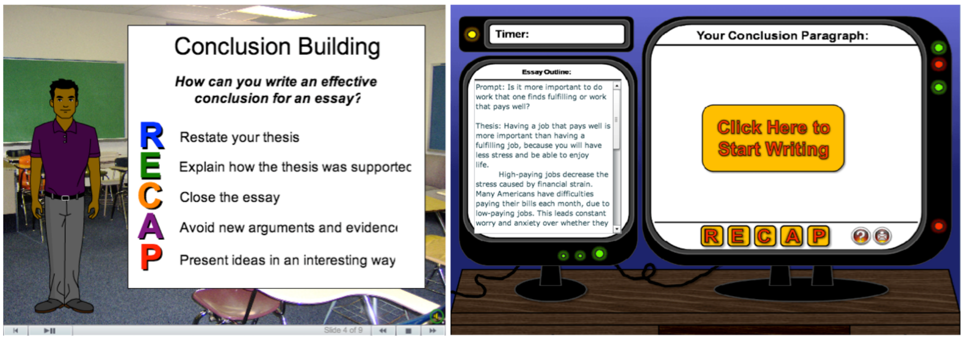
Grimes and Warschauer (2010) and Warschauer and Grimes (2008) have studied users' perceptions of AWE. Warschauer and Grimes (2008) surveyed principals, teachers, and students in four schools (6th to 12th grade) about their experiences with *Criterion* or *MyAccess*. Users reported that the systems improved students' writing motivation and essay quality, but the systems were not used often due to curriculum conflicts. For example, the systems could not support every writing genre that teachers needed to cover. In a later study, Grimes and Warschauer (2010) examined perceptions of *MyAccess* over three years in four middle schools. Teachers reported that *MyAccess* saved time, allowed them to focus on deeper concepts, and motivated students, but they also expressed doubts about scoring accuracy. Teachers also felt that *MyAccess* was more suitable for persuasive writing than teaching informative, narrative, or analytical writing. Likewise, teachers judged *MyAccess* to be helpful for teaching sentence fluency and conventions, but less helpful for ideas, organisation, voice, or word choice. Finally, students described *MyAccess* as usable and enjoyable, and stated that it increased their writing confidence and quantity. Unfortunately, students also had difficulty with understanding the feedback and managing the amount of feedback received. Some teachers had to develop supplemental handouts to assist students with implementing the feedback.

In sum, there appears to be a disconnect between scoring accuracy, system usability, and instructional efficacy. Although scoring processes are fairly accurate, broad instructional benefits are not well-established and many users remain skeptical. Further concerns relate to the validity of the scores (Clauser et al., 2002) and feedback. The linguistic features needed to accurately assign scores typically comprise only a subset of the characteristics that can describe an essay. Such parsimony is desirable for scoring algorithms, but strongly data-driven models may not capture the unique strengths and weaknesses of an essay. Consequently, the resulting feedback algorithms may overlook less common writing problems and may not map onto the kinds of formative feedback a writing instructor would offer. Across research studies, it appears that instructional efficacy and perceived utility may be improved to the extent that AWE can provide more individualised formative feedback.

3 Formative feedback in an intelligent tutoring system for writing

In contrast to AWE approaches for writing instruction, which emphasise intensive essay-based practice, intelligent tutoring systems seek to enhance writing development via explicit and adaptive strategy instruction and practice. For example, the *escribo* system offers structured scaffolding during the writing process along with informational hints (e.g., Proske et al., 2012). *Summary Street* (e.g., Franzke et al., 2005) is a LSA-based tutoring system that focuses on summary writing, including visual feedback on how well students' summaries cover the content of source texts.

Figure 1 Example of a (a) conclusion building lesson and (b) the lockdown practice game (see online version for colours)



Conclusion Building

How can you write an effective conclusion for an essay?

- **R** Restate your thesis
- **E** Explain how the thesis was supportec
- **C** Close the essay
- **A** Avoid new arguments and evidenc
- **P** Present ideas in an interesting way

Essay Outline:

Prompt: Is it more important to do work that one finds fulfilling or work that pays well?

Thesis: Having a job that pays well is more important than having a fulfilling job, because you will have less stress and be able to enjoy life.

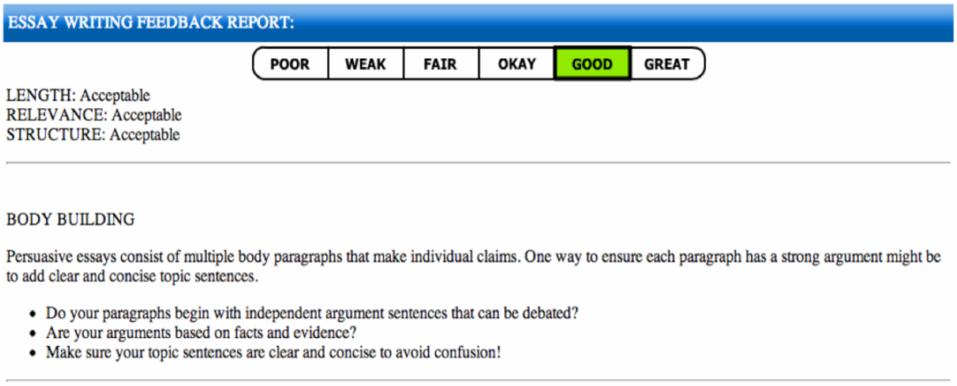
High-paying jobs decrease the stress caused by financial strain. Many Americans have difficulties paying their bills each month, due to low-paying jobs. This leads constant worry and anxiety over whether they

(a) (b)

The *Writing Pal* (W-Pal) is an intelligent tutoring system that provides strategy instruction, game-based practice, essay-based practice, and both summative and formative feedback for developing writers (Figure 1; McNamara et al., 2012; Roscoe and McNamara, 2013; Roscoe et al., in press). Eight modules offer strategies to aid students in completing various stages of the writing process, including prewriting (*freewriting* and *planning*), drafting (*introduction building*, *body building*, and *conclusion building*), and revising (*paraphrasing*, *cohesion building*, and *revising*). Each module includes *instructional videos* narrated by an animated character and *educational games* that enable strategy practice. In *identification* games, students examine short texts to identify strategy applications or examples. For example, in *undefined and mined*, players defuse bombs by identifying undefined referent words. In *generative* games, students write short texts while applying one or more strategies. For instance, in the conclusion building module *lockdown* game, players use strategies learned in the conclusion Building lessons [Figure 1(a)] to fend off computer hackers by writing conclusion paragraphs [Figure 1(b)].

Students also author practice essays that receive holistic scores and formative feedback driven by NLP algorithms. W-Pal utilises *Coh-Metrix* and related tools (e.g., McNamara et al., 2013; McNamara and Graesser, 2012) to analyse text on numerous dimensions, such as referential cohesion, connectives, lexical diversity, and LSA-based measures. Coh-Metrix also calculates syntactic complexity, measures psycholinguistic word data (e.g., concreteness and hypernymy), and assesses rhetorical features of the text (i.e., features of introduction and body paragraphs). A hierarchical series of statistical models use linguistic properties to determine essay quality and assign scores. Currently, essays are assigned a holistic rating from *poor* to *great* (six-point scale). Writers then receive formative feedback (see Figure 2) on writing goals and strategies via a series of algorithms on *legitimacy, length*, *relevance*, *structure*, *introduction*, *body*, *conclusion*, and *revising*. Unlike most AWE tools, W-Pal provides no feedback on low-level errors (e.g., spelling) and provides less feedback overall to avoid overwhelming users (Grimes and Warschauer, 2010). W-Pal gives one feedback message on one *initial topic* (i.e., the *first* problem detected in the series of checks). Students can voluntarily request more feedback (up to ten total messages) on that topic or on an additional *next topic* (i.e., the *next* problem detected).

Figure 2 Example of W-Pal feedback with a holistic rating and strategies for body building (see online version for colours)



LENGTH: Acceptable
RELEVANCE: Acceptable
STRUCTURE: Acceptable

BODY BUILDING

Persuasive essays consist of multiple body paragraphs that make individual claims. One way to ensure each paragraph has a strong argument might be to add clear and concise topic sentences.

- Do your paragraphs begin with independent argument sentences that can be debated?
- Are your arguments based on facts and evidence?
- Make sure your topic sentences are clear and concise to avoid confusion!

Formative feedback is a central component of W-Pal design because of the focus on strategy instruction. W-Pal algorithms must be sensitive to holistic essay quality and individual students' strategies, strengths, and weaknesses. Our efforts to design such algorithms have further revealed the inadequacy of feedback systems that are too strongly grounded within scoring systems. W-Pal demonstrates agreement between algorithm-assigned and human-assigned scores that is comparable to AWE systems (McNamara et al., 2013) but the translation of scoring algorithm output into pedagogically-valid feedback is a particular challenge. W-Pal feedback currently addresses only eight broad categories. Although a variety of feedback messages have been constructed for each topic, the long-term goal is to provide formative feedback recommendations that are much more specific, fine-grained, and individualised.

To achieve these goals, this paper explores an alternative method for developing feedback algorithms that are not derived from scoring algorithms. Instead, we first identify guidelines offered by established writing standards, rubrics, and style guides, and then map these guidelines onto computational linguistic measures. More specifically, we begin by identifying general pedagogical principles (e.g., 'maintain an objective tone') and then connect these principles to specific computational indices that may capture the construct. Subsequent analyses produce threshold values for the target variables that can be used to govern when and how to respond to problems in students' essays.

4 Method: pedagogically-guided algorithm development

4.1 Essay corpus

From prior research, we assembled a corpus of 939 essays written by high school or freshman college students. All essays were written in response to SAT-style persuasive essay prompts with a 25-minute time limit. Prompts addressed a variety of topics: choices (n = 79), competition (n = 126), fame (n = 133), fitting in (n = 35), heroes (n = 158), images (n = 126), memories (n = 45), optimism (n = 56), uniqueness (n = 155), and winning (n = 35). All essays had been previously scored by trained human raters (inter-rater reliability $r \geq .70$) using a SAT-based rubric (six-point scale).

4.2 Selection of linguistic indices

We first identified writing principles from diverse sources, such as the common core state standards (CCSS), writing style guides, and the SAT rubric. For example, the CCSS for writing (National Governors Association, 2010) delineate a variety of goals for persuasive writing. Students are expected to 'write arguments to support claims with clear reasons and relevant evidence', and to 'use words, phrases and clauses to clarify the relationships among claim(s), reasons, and evidence'. The standards also address essay tone, stating that students should 'establish and maintain a formal style'. Writing style guides similarly offer lessons 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. Finally, scoring rubrics used to assess writing are another source of pedagogical information (de la Paz, 2009). The SAT 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 'skilful 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" (Camara, 2003).

An initial set of linguistic indices was mapped onto selected writing guidelines by selecting the most explicit and transparently applicable measures available. For example, the principle of 'use larger words' was linked to the 'mean syllables per word' measure, which directly assesses word size and sophistication. Likewise, the principle of 'use words to link ideas' was mapped onto measures of connective phrase usage.

Importantly, the mappings explored in this study were not intended to be exhaustive or final. We examine these particular mappings between guidelines and indices as a first step in this endeavour. As we consider later in the discussion section, we also expect to further explore questions about optimising such mappings in future studies.

4.2.1 Vocabulary and language

Writing guidelines emphasise the use of sophisticated, concrete, and precise vocabulary to express ideas with clarity and specificity. Word size was assessed via the *mean number of syllables* per word in the essay. Word variation was assessed by *lexical diversity* (i.e., occurrence of unique words relative to essay length). *Word concreteness* assessed the extent to which words referred to concrete sensory experiences (e.g., apple) versus abstract concepts (e.g., truth). *Word hypernymy* assessed the mean level of word specificity in the text.

4.2.2 Organisation and structure

Writing guidelines specify that ideas should be expressed in well-structured, purposeful paragraphs with elaborated arguments and evidence. Basic paragraph structure was assessed by the *total number of paragraphs* in the essay. *Number of words* and the *average length of paragraphs* were used as rough measures of elaboration. *Conclusion paragraph n-grams* served to detect the presence of concluding statements or concluding sections within the essays.

4.2.3 Cohesion

Writers are often taught to use linking words and other techniques to build cohesion and communicate conceptual relationships. 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. The LSA *paragraph-to-paragraph* measure used LSA to assess global cohesion across paragraphs. *Argument overlap* captured cohesion at the level of sentences and brief sections of text by examining overlap in key words and phrases.

4.2.4 Objective and formal tone

Writers are typically directed to maintain an objective and formal tone, including factual and appropriate evidence. The *incidences of first, second, and third-person pronouns* were used to assess objectivity. Use of examples was assessed by the incidence of *exemplification n-grams*, such as 'for instance', which often communicate the inclusion of examples. The incidence of *academic words* (e.g., *analyse* and *integrate*) measured the use of formal language. The incidence of *date-time words* (e.g., *November*) indicated references to specific dates or time periods.

4.3 Establishing feedback thresholds

Because the writing process is complex, it can be difficult to identify specific problems within essays and determine whether these errors require feedback. One method is to establish algorithms such that only essays that fall above or below a strict threshold receive feedback. To build objective, quantitative thresholds, we employed a straightforward *binning process* that grouped essays into four categories based on mean and standard deviation values for each variable provided in Table 1. 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 *greater than one standard deviation below* the mean were placed in the 'lowest' bin. For example, the mean number of words in essays from this corpus was 312.6 (*SD* = 112.1). Essays containing more than 424 words were placed in the highest bin for number of words, and essays containing between 112 and 423 words were categorised in the high bin. Likewise, essays containing between 201 and 311 words were categorised in the low bin, and essays containing fewer than 200 words were placed in the lowest bin.

5 Results: algorithm validation

The initial validation process occurred in two phases. We first assessed whether holistic scores differed based on category. For example, did essays in the lowest bin for hypernymy earn lower scores, as would be expected by pedagogical guidelines that emphasise precise word usage? Second, we qualitatively analysed the degree to which example essays from the low-scoring bins exhibited the targeted problem. For instance, were essays in the lowest bin for hypernymy indeed more vague? This step is crucial for demonstrating construct validity for chosen indices.

5.1 Essay scores based on bin categorisation

Table 1 reports the average essay scores for each bin category. One-way ANOVAs revealed main effects of bin category for most of the linguistic variables. In many cases, essays in the lowest or highest bin earned a significantly lower score than essays in other bins. In a few instances, the pattern appeared to be curvilinear; both the highest and lowest bins scored poorly compared to the high and low bins. The following sections summarise the pattern of results for each examined variable. Due to space limitations, pair-wise comparisons between bins are not shown.

5.1.1 Vocabulary and language

Consistent with writing guidelines, essays that relied on short, abstract, and vague words earned lower scores. Similarly, essays that exhibited minimal variation or diversity in word choice also earned lower scores.

Table 1 Guidelines, measures, and mean essay scores by bin category

Writing guideline	*Linguistic measure*	*Bin category*				*F*
		1	*2*	*3*	*4*	
Vocabulary and language						
‘Use larger words’	Mean syllables/word	*2.5*	2.9	3.2	3.4	29.38[a]
‘Use varied words’	Lexical diversity	*2.4*	3.0	3.2	3.3	25.50[a]
‘Use concrete words’	Word concreteness	*2.5*	3.0	3.2	3.2	20.36[a]
‘Use precise language’	Word hypernymy	*2.4*	3.0	3.2	3.4	40.87[a]
Organisation and structure						
‘Develop a clear paragraph structure’	Number of paragraphs	*2.1*	2.8	3.3	3.4	130.07[a]
‘Provide sufficient elaboration’	Mean paragraph length	*1.8*	3.1	3.0	*2.5*	13.38[a]
	Number of words	*2.0*	2.9	3.3	3.8	135.67[a]
‘Provide a concluding statement’	Conclusion *n*-grams	3.0	3.1	3.1	*2.7*	6.26[a]
Cohesion						
‘Use words to link ideas’	Additive connectives	2.9	3.0	3.1	3.0	2.49
	Causal connectives	3.0	3.1	3.1	3.0	0.84
	Logical connectives	3.0	3.1	3.1	*2.7*	7.34[a]
	Negation connectives	3.0	3.1	3.1	*2.7*	7.69[a]
‘Create cohesion among ideas’	LSA paragraph-to-paragraph	*2.2*	3.2	3.2	3.2	54.47[a]
	Argument overlap	*2.9*	3.1	3.0	*2.9*	2.90[b]
Objective and formal tone						
‘Establish an objective tone’	1st person pronouns	-	3.1	3.0	3.0	0.60
	2nd person pronouns	-	3.2	2.7	2.4	41.63[a]
	3rd person pronouns	*2.8*	3.0	3.1	3.1	2.60
‘Provide examples’	Exemplification *n*-grams	*2.8*	3.3	3.1	*2.9*	17.01[a]
‘Maintain a formal style’	Academic words	*2.4*	3.1	3.2	3.3	27.13[a]
‘Refer to specific dates and times’	Date-time words	-	*2.9*	3.3	3.2	17.79[a]

Notes: 1 = lowest, 2 = low, 3 = high, 4 = highest. [a]p < .001; [b]p < .05.
Empty cells, indicated by a dash, represent cases in which no essays were categorised in a particular bin. Italics entries indicate bins that scored significantly lower than other bins.

5.1.2 Organisation and structure

Essays that were very short (i.e., fewer words) or contained fewer paragraphs (e.g., a single block of text) earned lower scores. In contrast, essays that contained a high frequency of conclusion *n*-grams (e.g., *in conclusion* and *in the end*) earned lower scores.

Conclusion *n*-grams signal that the author is presenting a summary of ideas or otherwise bringing an essay to a clear end. Overuse of such phrases may indicate that the essay has a muddled organisation or flow.

Paragraph length seemed to exhibit a curvilinear relationship with essay scores. Essays with very short paragraphs earned lower scores, perhaps because the ideas were presented in a terse and unelaborated manner. By contrast, essays with very long paragraphs may result from compressing too many ideas and arguments into one paragraph when it would have been more appropriate to discuss them separately.

5.1.3 Cohesion

In this study, no significant differences were observed for additive (e.g., *moreover*) or causal (e.g., *as a result*) connectives. However, essays that exhibited very frequent use of logical (e.g., *thus*) or negation (e.g., *in contrast*) connectives earned lower scores. A high density of logical connectives may indicate essays in which claims are stated but not supported. Likewise, negations may signal essays in which writers contradict themselves or fail to adopt a clear stance. Together, these results suggest that transition words may be a strategy that students should apply cautiously and precisely.

The argument overlap measure captures cohesion across sentences and demonstrated a curvilinear relationship with essay scores. Essays with very high overlap and repetition across sentences earned lower scores (i.e., very redundant text), as did essays in which key words, ideas, and themes were only rarely threaded across sentences. Finally, the LSA paragraph-to-paragraph measure assesses global cohesion across paragraphs in an essay. Essays with a very low LSA paragraph-to-paragraph value earned lower scores, indicating that lower global cohesion (i.e., conceptually disconnected paragraphs) negatively impacted scores.

5.1.4 Objective and formal tone

In this study, no significant differences were observed for first or third-person pronoun use. These measures likely failed to capture the nuanced ways that such perspectives are used well or poorly in an essay. For example, writers who can effectively use personal experiences to argue a point will tend to earn better scores, whereas writers who rely on unsupported opinion statements (e.g., I believe) may earn lower scores. In contrast, overuse of second person pronouns significantly lowered essay scores. The second-person perspective is highly informal and communicates a level of familiarity with the audience that may not be appropriate in academic writing.

Examples are a key means of presenting evidence in prompt-based essays. Interestingly, the pattern for exemplification *n*-grams was curvilinear. As expected, writers who rarely used exemplification markers received lower scores. However, overuse of these markers also negatively impacted scores, perhaps because students tended to overuse examples that were underdeveloped or irrelevant to the topic. Next, the incidence of academic words was associated with higher essay quality. These terms are extracted from diverse academic disciplines and thus relate to formal language appropriate to scientific and empirical discourse. Finally, in alignment with writing guidelines, writers who incorporated more references to specific dates and times also earned higher scores.

5.2 Qualitative analysis of low-scoring essays

The preceding analyses indicate that essays categorised based on many of the selected linguistic variables differed significantly in holistic scores. However, we cannot immediately assume that the variables captured the target writing guidelines. Thus, essays from low-scoring bins were read and hand-annotated to establish whether a meaningful writing problem was demonstrated. Subsequently, formative feedback recommendations can be developed to guide writers toward improvement. The following sections provide excerpts from *word hypernymy, LSA paragraph-to-paragraph, second-person pronouns*, and *exemplification*, which highlight a variety of different bin categories. Spelling and punctuation were not corrected; italics were added to highlight key phrases and examples.

5.2.1 Word hypernymy

Word hypernymy was intended to detect essays that used vague wording. In the excerpt below (from the lowest bin), the writer used many vague terms and failed to specify concepts. For example, the author uses the phrase ‘once in a while’ rather than providing details about when and why it is ‘good’ to be optimistic. Similarly, the writer indicates that realism is helpful when events are unexpected or undesirable but offers no specific examples of such cases. Formative feedback for this author could offer strategies for choosing more meaningful and precise terms, recognising undefined referents, and elaborating explanations.

> It is *good* to be realistic. It is *good* to be optomistic *once in a while*, but you should *usually* should be optomistic. Being realistic will help prepare you if the outcome is *not what you want it to be*. If you are always optomistic you will not be prepared if something does not *turn out the way you want it to*. If you are realistic you can think about *good and bad outcomes* to a situation.

A contrasting example is selected from the highest bin for word hypernymy, which demonstrates greater specificity in two ways. First, the author employs words with more precise and richer meaning, such as ‘demolished’, ‘reckless’, and ‘lavish’. Second, the author includes specific examples related media coverage related to celebrity divorces, deaths, and criminal activities. With regards to feedback, this writer could be praised for their use of more specific examples and encouraged to further develop this aspect of their writing.

> In today's society, views of heroes have *demolished* the true meanings of a role model. *Modern* day *artist*, *actors*, and even reality TV stars are *hailed* due to their *reckless antics* and *lavish spending*. This misconception of heroes can be blamed on the *media*, which uses most of their *coverage* time discussing widely *anticipated divorces*, such as *Tiger Woods* or *Tom Cruise*, deaths, like *Michael Jackson* or *Anna Nicole Smith*, or even *jail* time, including *T. I* and *Lil Wayne*.

5.2.2 Lexical diversity

Lexical diversity assessed whether essays demonstrated repetitive vocabulary. In the following excerpt (lowest bin), the writer repeats many words, rarely straying from the terms of ‘destiny, ‘achieve’, and ‘choice’. Feedback for this student could provide strategies for identifying and replacing repeated words with meaningful synonyms. At a deeper level, this student also appeared to struggle to develop the ideas that were presented, which may have been one underlying cause of the redundancy. Thus, additional formative feedback might suggest methods for generating ideas.

> *I agree with this quote*, "*Destiny* is not a matter of *chance*. It is a matter of *choice*. It is not a thing to be waited for, it is a thing to be *achieved*." This *quote* means that the more *choices* you make the closer you are to your *destiny*. It also means that you cant sit around and wait for your *destiny* you must *achieve* it.
>
> *Destiny* is determined by *chance* and by the *choices* you make. I believe that the *choices* we make, can lead to abilities and talents that show who we truly are. For example, you may be talented at football but if you do not choose to play your *destiny* will not be *achieved*.

An except from the highest bin for lexical diversity exhibits much more varied wording, and also shows a more sophisticated use of words, overall. Although lexical diversity does not necessarily guarantee that a writer employs an advanced vocabulary, those two characteristics of writing may cooccur for better writers.

> In 1492, Columbus sailed the ocean blue. This *popular* children's *rhyme represents* the much greater event that acted as a key *turning point* in history. Columbus' *voyage* brought *death* and *devastation* to the *indigenous* peoples of the Americas, yet it also *spread* new foods and people that would *revolutionize* the *global economy*. His *purpose* for *embarking* on such a *dangerous journey* was not personal *satisfaction*, it was none other than money. Columbus believed he was sailing to Asia, not an *undiscovered continent*. He simply wanted to find *spices* and other *valuable* resources he knew to be in Asia. Indeed, when he arrived, he *exploited* the land and its people in order to gain the most personal *profit*.

5.2.3 LSA paragraph-to-paragraph

This LSA-based measure evaluated aspects of global cohesion via semantic similarity across paragraphs. The excerpt below (lowest bin) shows two consecutive paragraphs from an essay. The paragraphs are neither brief nor lacking in detail; the writer shows knowledge of literature and current world events relevant to the topic (i.e., influence of impressions and appearances). However, the relationship of these paragraphs to an *overarching argument* is less clear. The author has presented two elaborated examples that he or she believes demonstrates an idea, but this idea is not made explicit (e.g., in a topic sentence). Consequently, the two paragraphs seem isolated from each other and lack global cohesion.

Feedback for this writer may not need to discuss idea generation or elaboration. This writer also appears to understand the value of specific and concrete examples. However, the writer may benefit from learning about cohesion-building strategies, such as restating key themes throughout an essay to maintaining conceptual flow.

> In the Tragedy of Othello by William Shakespeare, the protagonist, Othello, is misled by Iago. Iago gains Othello's trust, and uses it to destroy Othello. By appearing to be Othello's friend, Iago is able to manipulate Othello and bring him to his downfall. Throughout the play, Othello trusts Iago completely even as Iago turns Othello against his wife. Had Othello seen through Iago's charade of friendship, Othello could have evaided his tragic end. Iago's surface appearance effected Othello's judgement, so he couldn't see what was really happening.

> In the media recently, democratic representative Weiner has faced a scandal relating to an alleged lewd image he sent to a girl on Twitter. The image in question contains no nudity, but the media is in an uproar over the story, and Rep. Weiner's career may be at stake. This image has no real bearing on Rep. Weiner's ability to represent his state in congress; nor does it break any laws. His image, though, will be damaged irrepirably. Simply because he has made a bad impression, he might not be able to continue in congress.

In a contrasting case from the highest bin for LSA paragraph-to-paragraph, the author maintains a clear theme from the one body paragraph to another. Specifically, the author discusses how celebrities can be role models, especially for children, by supporting charitable campaigns and helping others.

> Although many celebrities today sometimes just want to act in movies and sell tabloids, some do support important campaigns. They appear and help run drives to raise money for important causes from cancer to abuse. The celebrities who help important causes can be heroes. They become important influences in other people's lives. Children who see the celebrities helping and are influenced to help since the people, the celebrities, they see on television, on the computer, and in magazine print are also helping. The heroes who help, support, and advertise campaigns for causes such as cancer and abuse should be admired because they influence those around them in good ways.
>
> Celebrities can also act as good influences when they exist as heroes through how they act. As many children today see celebrities splashed across tabloids doing drugs and in jail, those celebrities' actions may influence children into believing drugs and jail time are okay and typical. However, celebrities who are viewed by children helping the common people around them are heroes. Celebrities are just like normal people in the fact they have the same choice of whether or not they actually help other people who need help. The celebrities who take time to help people in need and show they are helping through their actions should be admired because they are real heroes.

5.2.4 Second-person pronouns

Frequent use of the second-person perspective is associated with informal communication and is atypical of academic writing. In this excerpt (highest bin), the tone is indeed informal. Another trend noted for essays in this bin was the frequency of speculative if-then claims, such as "if you are trying to employee of the month at your job, you would do whatever it takes". Rather than building a case to support their claims, writers made a series of pronouncements directed at the reader. Feedback could help this writer use other perspectives to establish an objective stance. Further exploration may show that writers who abuse second-purpose pronouns would also benefit from feedback on avoiding speculative arguments or claims.

> For example if *you* are trying to employee of the month at *your* job, *you* would do whatever it takes to get it. *You* would be competing against the other co-workers. Competition gets *you* more focused to whatever *you* want to achieve, if *your* unfocused it may not work out. It makes *you* want to try harder to be the best, kind of like pressuring *you*. Competition is like a motive, it keeps pushing to do whatever it takes until *you* get it.

By contrast, the following example from the low bin for second-person pronouns demonstrates a somewhat more objective and formal tone by avoiding the use of 'you' and 'your' pronouns.

> There are many individuals today that would like to create something new, and yet, it seems that seeing brand new ideas come about is becoming more infrequent. However, this is not to say that people cannot truly be original. While spawning a brand new idea is always a possibility, people could also take a previously invented thought and improve upon it with their own originality. Such an abundance of new tools and knowledge is available to be used to update previous ideas that have already been done to today's standards.

5.2.5 Exemplification n-gram*s*

These *n*-grams assessed whether writers used examples. Analyses showed that *both* the lowest and highest bins for this variable received lower scores. In the first excerpt (lowest bin), no exemplification was observed; the essay did not offer any specific examples. For instance, the author does not describe cases of human jealousy or ambition, nor does the author describe people who are 'unable to change'. Feedback for this writer might suggest specific methods for choosing and developing examples, such as freewriting to generate ideas, or drawing upon world knowledge gained from the news or schoolwork.

> The human kind is a very jealous and ambitious 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.

In the excerpt below (highest bin), examples are mentioned but not elaborated with pertinent details. Feedback for this student could discuss different forms of evidence that can be used to strengthen an essay, such as objective facts and relevant anecdotes.

> 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.

6 Discussion

Valid formative feedback is a key element of instructional technologies designed to support writing development. Growth in writing depends upon the clear communication of writing strategies, criteria, and goals. Thus, feedback that is invalid, overwhelming, or vague may explain why some AWE systems show high scoring accuracy but uncertain instructional efficacy. In this study, we described a method for creating pedagogically-guided feedback algorithms. Rather than grounding feedback algorithms in scoring algorithms, feedback thresholds were created by linking writing principles from established standards to quantifiable linguistic indices.

In a two-step validation process, many of the algorithm-based categorisations appeared to capture meaningful distinctions in essay quality. In most cases, essays at the extreme end of the continuum for a given variable showed significantly lower scores than essays in other bins. For example, essays in the lowest bin for word hypernymy (i.e., average word hypernymy was more than one standard deviation below the corpus mean) earned lower scores than essays categorised in the high and highest bins. Subsequent qualitative analyses offered insights into problems that individual writers may exhibit and ways to address these problems via formative feedback. For instance, essays in the lowest bin for hypernymy indeed exhibited use of more vague and imprecise wording and examples. Thus, the results offer supportive evidence that this pedagogically-guided algorithm methodology produces meaningful and useful information for delivering feedback. When these threshold algorithms are incorporated into future versions of W-Pal, additional formative feedback messages can be authored to provide actionable strategy suggestions, such as the example below:

> The use of more precise and meaningful words can help an author express his or her ideas more clearly and avoid confusing the reader.
>
> - Try to use words that express *exactly* what you are thinking.
> - For example, instead of using the word ‘dog’, try writing about the specific type of dog, such as ‘poodle’ or ‘golden retriever’.
> - A thesaurus is a great way to learn more specific words, but always make sure to double-check the meaning in a dictionary!

Another feedback message based on the hypernymy binning algorithm could offer strategies for elaborating one’s examples and evidence:

> Specific examples provide evidence that can convince readers to support the main arguments and claims in an essay.
>
> - Try to provide several examples to support every argument.
> - For each major example, provide a few sentences that explain the example and how it supports your ideas.
> - When revising, look for examples that are described very briefly and give the reader more specific details.

An important contribution of this approach is that it may enable educators and developers to more confidently target problems in individual essays. Essays classified in different bins may differ by a standard deviation or more on the target variable, which helps to ensure that essays in low-scoring bins truly merit feedback on the target issue. Likewise, the same thresholds could be used to guide individualised positive feedback for essays in high-scoring bins. For W-Pal, this approach may allow developers to refine broad feedback categories (e.g., ‘body building’) to offer more specific and individualised features, such as the use of certain types of evidence (e.g., references to specific dates), objectivity (e.g., formal language and pronoun use), and logical flow across sentences (e.g., argument overlap). However, as more feedback categories are enabled, it remains important to not overwhelm users. Balancing the amount of feedback *delivered* against the range of feedback *possible* remains an important design question for AWE and intelligent tutoring systems (Roscoe et al., 2012).

6.1 Directions for future research

The results presented here represent an initial step in exploring the potential value of this approach. To further validate this approach, the threshold algorithms developed here will need to be implemented in W-Pal (or a similar system) and the instructional benefits of receiving such feedback must be empirically evaluated. In addition, there remain several questions to be addressed in future work. First, the mappings between writing guidelines and linguistic indices were one-to-one. Each principle was associated with only one linguistic measure. However, it is plausible that some essay qualities may best be represented by a cluster of variables. In these data, we observed that essays sometimes exhibited comorbidity among essay problems. For example, writers who relied on second-person perspective often also put forth speculative arguments that were not supported by evidence. Such writers may have difficulties with adjusting their writing to different audiences, instead defaulting to a conversational style used with friends. Likewise, some essays demonstrated cooccurrence of positive writing attributes, such as high lexical diversity and use of advanced vocabulary.

In future research, it may be important to consider how essays display clusters of categorisations across variables. Categorisations that often cooccur might be combined into a single pedagogical principle or algorithm. For example, if an essay is placed in the lowest bin for lexical diversity (i.e., very repetitive wording) *and* the lowest bin for hypernymy (i.e., very vague wording), then this might signal a deeper, underlying problem. As an alternative to providing feedback recommendations for each separate issue, the student may benefit from feedback that addresses more fundamental strategies or instruction related to overall word choice or vocabulary. Thus, in addition to performance on specific measures, the array of concerns revealed within individual students' work may be a powerful source of information for intelligent tutoring systems or classroom teachers to give personalised and adaptive formative feedback.

A second question relates to the manner in which linguistic data are obtained for essays. In this research, *Coh-Metrix* was used to analyse the texts. However, one strength of the approach described here is that it can be easily applied to other automated textual analysis tools. As one example, *Linguistic Inquiry and Word Count* (LIWC) is a text analysis tool that uses categorical word dictionaries to evaluate thematic and rhetorical language use (Pennebaker et al., 2007). The dictionaries are hierarchical and consist of word lists related to particular concept or theme (e.g., 'family' words and 'occupational' words). LIWC contains around 4,500 words and word stems across several dictionaries. When analysing text, LIWC provides a count of the words that occur within a dictionary, and these count values can be used to separate essays into bins based on means and standard deviations. Pedagogically, writers are discouraged from relying heavily on 'emotional appeals' in lieu of factual evidence. If students' 'emotion words' count is used to represent that pedagogical principle, then one might hypothesise that essays in the highest bin for emotional wording would earn a lower score. Such essays might benefit from feedback about objective language.

In this paper, we considered only a small subset of the writing principles that are commonly taught to students. This limitation was necessary in order to reasonably explore the new methodology. In practice, an instructional technology for writing need not be constrained to only a few features. Feedback tools may access comprehensive data about essays from multiple sources (i.e., combining multiple text analysis systems), thus providing a rich pool of topics and concerns that could be addressed via formative feedback. As just one example, Varner and colleagues (Varner et al., 2013) have demonstrated how automated textual analyses combining Coh-Metrix and LIWC can reveal misalignments in teachers' and students' criteria for writing.

A final concern is the extent to which separate algorithms may need to be established for different student populations. In this study, thresholds were developed based on a corpus comprising essays written by high school and college students on a variety of topics. On the one hand, such breadth may support relatively stable or generalisable thresholds that apply, with reasonable accuracy, to writers from different backgrounds. However, it is plausible that criteria for 'good' and 'poor' writing and formative feedback will differ based on the writer's academic grade level (Crossley et al., 2011), degree of proficiency with the written language (Silva, 1993), or learning disabilities (Graham and Harris, 1993). In practice, writers at different levels of age, development, or language fluency (e.g., English language learners versus native speakers of English) will likely exhibit different linguistic and textual patterns. What constitutes a highest or lowest bin essay for one population may not be equivalent in other populations with different backgrounds or needs.

7 Conclusions

The methods presented here have potential for efficiently generating flexible algorithms for feedback in computer-based writing instruction. Ultimately, we expect that such algorithms will have intrinsic pedagogical validity and contribute to greater effectiveness for improving students' writing proficiency. Nonetheless, a significant amount of research remains to be conducted on how feedback is developed and delivered through AWE and intelligent tutoring systems for writing. Through the work presented here, our overarching aim is to inspire and encourage further research, publication, and discussion in this area.

Acknowledgements

A preliminary version of this work was originally published in the proceedings of the 25th Annual Florida Artificial Intelligence Research Society (FLAIRS) conference (Roscoe et al., 2012). The current paper replicates that study using a substantially larger corpus of essays and expanded analyses. The research reported here was supported by the Institute of Education Sciences, US Department of Education, through Grant R305A080589 to Arizona State University. The opinions expressed are those of the authors and do not represent views of the Institute or the US Department of Education

References

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

Attali, Y. and Burstein, J. (2006) 'Automated essay scoring with e-rater V.2', *Journal of Technology, Learning, and Assessment*, Vol. 4 No. 3 [online] http://www.jtla.org (accessed 5 November 2010).

Burstein, J., Chodorow, M. and Leacock, C. (2004) 'Automated essay evaluation: the criterion online writing system', *AI Magazine*, Vol. 25, pp.27–36.

Camara, W.J. (2003) *Scoring the Essay on the SAT Writing Section*, College Entrance Examination Board, New York.

Clauser, B., Kane, M. and Swanson, D. (2002) 'Validity issues for computer-based tests scored with computer-automated scoring systems', *Applied Measurement in Education*, Vol. 15, No. 4, pp.413–432.

Crossley, S., Roscoe, R.D., Graesser, A. and McNamara, D.S. (2011) 'Predicting human scores of essay quality using computational indices of linguistic and textual features', *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, pp.438–440, AIED, Auckland, New Zealand.

Crossley, S., Weston, J., McLain Sullivan, S. and McNamara, D.S. (2011). 'The development of writing proficiency as a function of grade level: a linguistic analysis', *Written Communication*, Vol. 28, No. 3, pp.282–311.

de la Paz, S. and Graham, S. (2002) 'Explicitly teaching strategies, skills, and knowledge: writing instruction in middle school classrooms', *Journal of Educational Psychology*, Vol. 94, No. 4, pp.687–698.

Dikli, S. (2006) 'An overview of automated scoring of essays', *Journal of Technology, Learning, and Assessment*, Vol. 5, No. 1, [online] http://www.jtla.org (accessed 4 August 2009).

Graesser, A. C., and McNamara, D. S. (2012) 'Automated analysis of essays and open-ended verbal responses', in H. Cooper, P. Camic, R. Gonzalez, D. Long and A. Panter (Eds.): *APA Handbook of Research Methods in Psychology: Foundations, Planning, Measures, and Psychometrics*, American Psychological Association, Washington, DC.

Graham, S. and Harris, K. (1993) 'Self-regulated strategy development: helping students with learning problems develop as writers', *The Elementary School Journal*, Vol. 94, No. 2, pp.169–181.

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*, Vol. 8, No. 6, pp.4–43.

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

Kellogg, R., Whiteford, A. and Quinlan, T. (2010) 'Does automated feedback help students learn to write?', *Journal of Educational Computing Research*, Vol. 42, No. 2, pp.173–196.

Landauer, T.K., Laham, D. and Foltz, P.W. (2003) 'Automated scoring and annotation of essays with the Intelligent Essay Assessor', *Automated Essay Scoring: A Cross-Disciplinary Perspective*, pp.87–112, Erlbaum, Mahwah, NJ.

McGarrell, H. and Verbeem, J. (2007) 'Motivating revision of drafts through formative feedback', *ELT Journal*, Vol. 61, pp.228–236.

McNamara, D.S. and Graesser, A. (2012) 'Coh-Metrix: An automated tool for theoretical and applied natural language processing', in McCarthy, P.M. and Boonthum, C. (Eds.): *Applied Natural Language Processing and Content Analysis: Identification, Investigation, and Resolution*, pp.188–205, IGI Global, Hershey, PA.

McNamara, D.S., Crossley, S. and Roscoe, R.D. (2013) 'Natural language processing in an intelligent writing strategy tutoring system', *Behavior Research Methods*, Vol. 45, No. 2, pp.499–515.

McNamara, D.S., Raine, R., Roscoe, R.D., 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. and Graesser, A. (2012) 'The writing-pal: natural language algorithms to support intelligent tutoring on writing strategies', in McCarthy, P.M. and Boonthum, C. (Eds.): *Applied Natural Language Processing and Content Analysis: Identification, Investigation, and Resolution*, pp.298–311, IGI Globa, Hershey, PA.

National Governors Association Center for Best Practices and Council of Chief State School Officers (2010*) Common Core State Standards for Writing. National Governors Association Center for Best Practices*, Council of Chief State School Officers, Washington DC.

Pennebaker, J.W., Booth, R.J. and Francis, M.E. (2007) *Linguistic Inquiry and Word Count: LIWC*, [computer software], Austin, TX.

Proske, A., Narciss, S. and McNamara, D.S. (2012). 'Computer-based scaffolds to facilitate students' development of expertise in academic writing', *Journal of Research in Reading*, Vol. 35, No. 2, pp.136–152.

Roscoe, R.D. and McNamara, D.S. (2013) 'Writing Pal: feasibility of an intelligent writing strategy tutor in the high school classroom', *Journal of Educational Psychology*, Vol. 105, No. 4, pp.1010–1025.

Roscoe, R.D., Kugler, D., Crossley, S., Weston, J. and McNamara, D.S. (2012) 'Developing pedagogically-guided threshold algorithms for intelligent automated essay feedback', *Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference*, pp.466–471, AAAI Press, Menlo Park, CA.

Roscoe, R.D., Varner, L., Cai, Z., Weston, J., Crossley, S. and McNamara, D. (2011) 'Internal usability testing of automated essay feedback in an intelligent writing tutor', *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference*, pp.543–548, AAAI Press, Menlo Park, CA.

Roscoe, R.D., Varner, L., Weston, J., Crossley, S. and McNamara, D.S. (in press) 'The Writing Pal intelligent tutoring system: usability testing and development', *Computers and Composition*.

Rudner, L., Garcia, V. and Welch, C. (2006) 'An evaluation of the IntelliMetric essay scoring system', *Journal of Technology, Learning, and Assessment*, Vol. 4, No. 4, pp.3–21.

Shermis, M. and Burstein, J. (Eds.). (2003) *Automated Essay Scoring: A Cross-disciplinary Perspective*, Erlbaum, Mahwah, NJ.

Shermis, M., Burstein, J. and Bliss, L. (2004) 'The impact of automated essay scoring on high stakes writing assessments', Paper presented at the *Annual Meeting of the National Council on Measurement in Education*, San Diego, CA.

Shute, V. (2008) 'Focus on formative feedback', *Review of Educational Research*, Vol. 78, No. 1, pp.153–189.

Silva, T. (1993) 'Toward an understanding of the distinct nature of L2 writing: the ESL research and its implications', *TESOL Quarterly*, Vol. 27, No. 4, pp.657–677.

Sommers, N. (1982) 'Responding to student writing', *College Composition and Communication*, Vol. 33, No. 2, pp.148–156.

Varner, L., Roscoe, R.D. and McNamara, D.S (2013) 'Evaluative misalignment of 10th-grade student and teacher criteria for essay quality: an automated textual analysis', *Journal of Writing Research*, Vol. 5, No. 1, pp.35–59.

Warschauer, M. and Grimes, D. (2008) 'Automated writing assessment in the classroom', *Pedagogies: An International Journal*, Vol. 3, No. 1, pp.22–36.

Warschauer, M. and Ware, P. (2006) 'Automated writing evaluation: defining the classroom research agenda', *Language Teaching Research*, Vol. 10, No. 2, pp.1–24.