Tech., Inst., Cognition and Learning, Vol. 10, pp. 59–79
Reprints available directly from the publisher
Photocopying permitted by license only

Automated Detection of Essay Revising Patterns: Applications for Intelligent Feedback in a Writing Tutor

ROD D. ROSCOE[1], ERICA L. SNOW[2],
LAURA K. ALLEN[2] AND DANIELLE S. MCNAMARA[2]

[1]*Human Systems Engineering, Polytechnic School, Ira A. Fulton Schools of Engineering, Arizona State University, 150D Santa Catalina Hall, 7271 E. Sonoran Arroyo Mall. Mesa, AZ 85212, USA*
[2]*Learning Sciences Institute and Department of Psychology, Arizona State University, PO Box 872111, Tempe, AZ 85287, USA*

The Writing Pal is an intelligent tutoring system designed to support writing proficiency and strategy acquisition for adolescent writers. A fundamental aspect of the instructional model is automated formative feedback that provides concrete information and strategies oriented toward student improvement. In this paper, the authors explore computational methods for expanding automated feedback by taking into account students' essay revising patterns. High school students ($n = 87$) used the Writing Pal to write and revise a persuasive essay each day across eight daily sessions. Analyses of the linguistic properties of original and revised essays revealed that students were more likely to implement document-level revisions focused on improving elaboration, organization, and cohesion, rather than surface word-level edits focused on incorporating bigger words or less common words. Implications and applications of essay revising data for automated feedback and essay scoring are discussed.

Keywords: intelligent tutoring systems; writing; revising; formative feedback; strategy instruction; automated essay scoring; automated writing evaluation; computational linguistics

Computer-based tools for automated writing evaluation (AWE) are an increasingly popular means of delivering writing instruction (Shermis & Burstein, 2013). Established programs such as *Criterion* (e.g., Attali & Burstein, 2006; Burstein,

*Corresponding author: rod.roscoe@asu.edu

Chodorow, & Leacock, 2004) and *WriteToLearn* (e.g., Landauer, Laham, & Foltz, 2003) have been used by large numbers of students and teachers to practice and improve writing proficiency, and newcomers such as *LightSide* (Mayfield & Rosé, 2013) are striving to contribute innovative analytical methods.

The broad adoption of AWE tools has been based largely upon their success in scoring essays accurately and quickly. AWE programs generally yield holistic essay scores that are highly similar to those provided by human judges and can do so in seconds or minutes (Dikli, 2006; McNamara, Crossley, & Roscoe, 2013; Shermis & Burstein, 2013). The underlying scoring algorithms that drive these systems can be developed using a variety of methods, such as multiple regression models that predict essay scores using sets of linguistic features, Latent Semantic Analysis comparisons between student essays and benchmark corpora, and machine learning techniques to extract factors that discriminate among higher and lower quality essays (Dikli, 2006; Shermis & Burstein, 2013). A further motivation for AWE is that these programs also offer instructional support in the form of writing feedback (e.g., Grimes & Warschauer, 2010). Students who submit their essays to a system typically receive a holistic score along with summative feedback on errors in grammar, usage, structure, or key discourse elements (e.g., Attali & Burstein, 2006). Studies exploring the benefits of summative error feedback have found that student writers show improvements in terms of writing mechanics but not necessarily overall essay scores (e.g., Kellogg, Whiteford, & Quinlan, 2010; Rock, 2007). Overall, the field of automated writing evaluation has demonstrated remarkable power in using data extracted from student essays to inform meaningful scores, but more research is needed to develop effective instructional interventions based on these tools.

Our work on computer-based writing instruction has integrated automated writing evaluation and writing pedagogy from the earliest stages of design. The Writing Pal (W-Pal) is an intelligent tutoring system designed to support writing proficiency and strategy acquisition for adolescent writers (McNamara et al., 2012; Roscoe & McNamara, 2013; Roscoe, Varner, Weston, Crossley, & McNamara, 2014). W-Pal provides strategy instruction via lesson videos in which animated characters explain and demonstrate strategies pertaining to the broad writing process (i.e., prewriting, drafting, and revising). Viewing these lessons and completing quiz-like checkpoints unlocks educational games that enable targeted strategy practice (Roscoe, Brandon, Snow, & McNamara, 2013). Importantly, W-Pal also incorporates opportunities for students to practice authoring and revising prompt-based essays. Students can select from a variety of argumentative essay prompts and compose an essay that states and defends their viewpoint. The holistic quality of these essays is assessed via natural language

processing algorithms that consider lexical, syntactic, semantic, and rhetorical features of text, and are developed using a combination of analytic methods (McNamara et al., 2013; McNamara, Crossley, Roscoe, Varner, & Dai, 2015). These algorithms generate an essay score on a 6-point scale (from "Poor" to "Great"), similar to the SAT scoring rubric (Camara, 2003).

A foundational aspect of the instructional model for W-Pal is *automated formative feedback* (Roscoe & McNamara, 2013). Formative feedback offers concrete guidance and methods for student *improvement* (Shute, 2008), such as strategies for writing thesis statements, threading ideas to improve cohesion, or adopting an objective tone. This form of feedback is contrasted with summative feedback that evaluates *performance*, such as teacher-assigned grades or critiques of mechanical and organizational errors (Attali & Burstein, 2006). Although both modes of feedback are useful, research on writing instruction has emphasized individualized, formative feedback as essential to students' writing development because it communicates the knowledge and methods necessary for improvement (McGarrell & Verbeem, 2007; Sommers, 1982). In W-Pal, formative feedback is implemented via a hierarchical series of natural language algorithms assessing text *legitimacy*, *length*, *relevance*, *structure*, *introduction* quality, *body* quality, and *conclusion* quality. W-Pal automatically provides one feedback message on an *Initial Topic* (i.e., the *first* problem detected in the hierarchy). Subsequently, students can voluntarily request more feedback on that topic or on one additional *Next Topic* (i.e., the *next* problem detected). Thus, unlike many automated writing evaluation systems, W-Pal provides no feedback on low-level errors and provides less feedback overall to avoid overwhelming users (see Grimes & Warschauer, 2010). Figure 1 provides an example formative feedback report that offers strategies for body building and conclusion building.

In this paper, we explore methods for enhancing formative feedback in W-Pal by taking into account students' *essay revising patterns*. Writing is a complex process entailing iterative and recursive planning, drafting, and revising stages (Deane et al., 2008; Flower & Hayes, 1981). Planning involves the development and organization of ideas before writing, and drafting translates writers' plans and goals into a coherent text that communicates main ideas. Revising is defined as the refinement of a text to better achieve writers' goals. Skilled writers engage in more substantive revising to improve rhetorical strength (e.g., well-supported arguments), idea organization, and meaning, which is more likely to improve essay quality (Fitzgerald, 1987). These writers are more likely to edit their text in ways that improve the document as a whole, making it more cohesive and persuasive. For example, skilled writers may revise to develop a more logical flow of ideas such that new ideas clearly build from previously given ideas. Similarly,

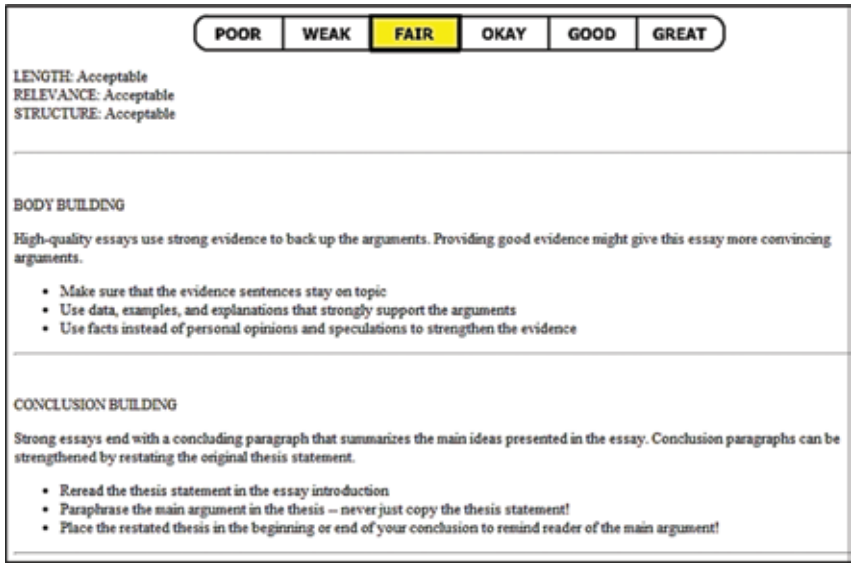

FIGURE 1
Example Essay Writing Feedback Report with recommendations regarding body building and conclusion building strategies.

these writers may revise to elaborate their ideas more substantively or provide additional supporting evidence. However, many students ignore revising or make only superficial edits to address spelling, grammar, and mechanics (Bridwell, 1980; Crawford, Lloyd, & Knoth, 2008; Fitzgerald, 1987; Sommers, 1980). At this level, for example, students might simply replace common words with longer, less frequent words. Bridwell (1980) analyzed Grade 12 students' revisions at seven levels: surface, words, phrases, clauses, sentences, multi-sentence, and text level (e.g., audience considerations). All students attempted to revise, but the most common revisions occurred at the word (31.2%) and surface levels (24.8%). Students revised mainly by improving their wording and correcting spelling, grammar, and punctuation errors. Similarly, Crawford and colleagues (2008) examined Grade 5 and Grade 8 students' revisions. Word level (∼40%), phrase level (∼25%), and punctuation (∼20%) revisions were again the most common, although these edits contributed to moderate improvements in quality.

A potentially powerful means of fostering writing development may be to provide formative feedback on both the discrete products of writing (i.e., original and revised drafts) and the *transformation* of writing across drafts. Students may benefit from specific formative feedback that addresses their revising process and

helps them enact substantive revisions. Whereas prior research on revising has utilized detailed human coding of essays (e.g., Bridwell, 1980), the aim of W-Pal is to be able to conduct meaningful analyses and provide feedback in an automated fashion. A necessary first step in this process is to examine student essay data and establish whether computational linguistic tools can meaningfully detect essay revisions. To this purpose, we analyze a corpus of essays written by high school students who interacted with W-Pal. We use Coh-Metrix and related tools (e.g., McNamara, Crossley, & Roscoe; 2013; McNamara & Graesser, 2012; McNamara, Graesser, McCarthy, & Cai, 2014) to assess how students' original drafts changed during revision and whether such revisions targeted shallow, word-level features or more holistic, document-level properties of the essay.

METHOD

Corpus

Essay data were obtained from an experiment conducted with W-Pal to contrast two modes of computer-based writing instruction and practice: a typical approach that emphasizes writing practice (i.e., writing and revising many essays with feedback) versus an alternative approach in which learners engage in less writing practice (i.e., fewer essays) but receive direct strategy instruction and game-based practice. Prior analyses of these data have indicated that the instructional condition that included direct strategy instruction fostered strategy acquisition, writing self-efficacy, and aspects of substantive revising (Roscoe et al., 2013; Roscoe, Snow, & McNamara, 2013), but both approaches supported improvements in writing quality and cohesion (Crossley, Roscoe, & McNamara, 2013; Crossley, Varner, Roscoe, & McNamara, 2013). The current work does not further examine these experimental effects and thus data from the two conditions are collapsed.

High school students ($n = 87$) from an urban area in the southwest United States participated in a 10-session summer program using W-Pal. The average age of students was 15.6 years, with 62.1% females. Ethnically, 5.7% of students identified as African-American, 12.5% as Asian, 19.3% as Caucasian, and 54.5% as Hispanic. Average grade level was 10.4 with 40.2% of students reporting a GPA of 3.0 or below. Most students self-identified as native English speakers ($n = 49$), although many self-identified as English Language Learners (ELL, $n = 38$).

The program was conducted by the researchers in a classroom-like, laboratory setting. The researchers provided no writing instruction or tutoring to the students, and students were monitored to ensure that they worked individually rather than collaboratively. Students began each of eight sessions by writing a prompt-based

essay. Students were allotted 25 minutes to draft their essay and 10 minutes to revise after receiving feedback. A different argumentative prompt was assigned each day in the following order: *Planning*, *Winning*, *Patience*, *Heroes*, *Perfection*, *Uniformity*, *Beliefs*, and *Fame* (Table 1). Each prompt was purposefully modeled after the kinds of timed writing assignments often included on standardized tests (e.g., the SAT exam). Students were asked to read a statement about a topic and then "Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations." An example prompt on the topic of *Uniformity* is given below:

> From the time people are very young, they are urged to get along with others, to try to "fit in." Indeed, people are often rewarded for being agreeable and obedient. But this approach is misguided because it promotes uniformity

TABLE 1
Brief summaries of persuasive writing prompts and questions

Prompt	**Summary**
Beliefs	Argue whether the veracity of an idea is based on whether many people believe it to be true. Are widely held views often wrong, or are such views more likely to be correct?
Fame	Argue whether people are more driven to succeed by external rewards or by internal rewards. Are people motivated to achieve by personal satisfaction rather than by money or fame?
Heroes	Argue whether famous or popular individuals deserve as much public respect, or deserve to be role models, as those have demonstrated heroic deeds. Should we admire heroes by not celebrities?
Patience	Argue whether individuals should seek instant results and gratification or should be patient about their goals. Is it better for people to act quickly and expect quick responses from others rather than to wait patiently for what they want?
Perfection	Argue whether being perfectionist or detail-oriented helps to achieve success or is a potential hindrance. Do people put too much importance on getting every detail right on a project or task?
Planning	Argue whether individuals should think seriously and plan ahead even when it is challenging or uncomfortable. Does every individual have an obligation to think seriously about important matters, even when doing so may be difficult?
Uniformity	Argue whether it is more beneficial for individuals to conform and follow the rules compared to embracing diverse opinions and perspectives. Is it more valuable for people to fit in than to be unique and different?
Winning	Argue whether individuals or society place an appropriate value on success and victory or if awards and prizes distract us from more important qualities. Do people place too much emphasis on winning?

instead of encouraging people to be unique and different. Differences among people give each of us greater perspective and allow us to make better judgments. Is it more valuable for people to fit in than to be unique or different?

Students authored a total of 696 essays but technical issues prevented several essays from being recovered. The final corpus contained 688 original drafts along with their revised drafts.

Linguistic Properties of Essay Revisions

In this study, we focused on automated detection of revising patterns at two levels: the *word level* and the *document level* (see Table 2). Essays were analyzed using a computational linguistic tool called Coh-Metrix. Coh-Metrix provides automated analyses of text along a variety of dimensions, ranging from basic text properties (e.g., number of words and sentences) to more sophisticated measures of text difficulty and complexity (e.g., narrativity and cohesion; for a more thorough description of Coh-Metrix indices, see McNamara et al., 2014). Table 2 summarizes the specific linguistic variables examined at each level. Word-level measures were related to students' word choice and vocabulary, such as word

TABLE 2
Word-level and document-level linguistic measures of revising

Linguistic Measure	Brief Description
Word-level	
Average syllables per word	Mean number of syllables for all words in the text
Lexical diversity	Incidence of unique words relative to the whole text
Word frequency	WordNet ratings of word frequency
Word concreteness	WordNet ratings of word concreteness
Word familiarity	CELEX measure of word familiarity
Polysemy	WordNet measure of word ambiguity (number of senses for a given word)
Hypernymy	WordNet measure of word specificity
First person pronouns	Incidence of pronouns such as "I" and "me"
Second person pronouns	Incidence of pronouns such as "you" and "yours"
Third person pronouns	Incidence of pronouns such as "he" and "they"
Document-level	
Number of words	Total number of words
Number of paragraphs	Total number of paragraphs
LSA adjacent sentences	Mean LSA cosine for adjacent sentences
LSA all sentences	Mean LSA cosine for all sentences
LSA paragraph-to-paragraph	Mean LSA cosine for paragraph to paragraph
LSA given/new	LSA measure of the amount of new information provided by each sentence
Narrativity component	Component score of text narrativity
Referential cohesion component	Component score of text referential cohesion
Deep cohesion component	Component score of text deep cohesion

frequency, concreteness, and hypernymy. For instance, word frequency measures assess how often a term appears in typical discourse, and hypernymy captures whether terms have more vague or precise meanings. Document-level indices captured revisions related to factors such as overall essay cohesion and structure. For example, the number of paragraphs in an essay can reveal key aspects of organization. Essays with only one or two paragraphs might also lack clearly defined sections (e.g., introduction, body, or conclusion). Similarly, Latent Semantic Analysis (LSA, Landauer, McNamara, Dennis, & Kintsch, 2007) was used to assess the semantic relatedness of paragraph and sentences within students' essays. The degree to which similar ideas, themes, and meaning are maintained across essay sections can be an important cohesion cue.

RESULTS

Writing Time and Essay Quality

Duration

Essays were initially examined with regard to the amount of time students spent writing (i.e., duration). These analyses considered only students who provided a complete set of timing data for all prompts (i.e., $n = 78$ for original drafts and $n = 76$ for revised drafts).

The time students spent writing was analyzed in a repeated-measures ANOVA with prompt as the within-subjects variable (see Table 2). Students spent approximately 21 minutes drafting their original essays, with significant differences across prompts over time. There was a significant linear effect, $F(1,77) = 14.44$, $p<.001$, a marginally significant quadratic effect, $F(1,77) = 3.93$, $p = .051$, and a marginally significant cubic effect, $F(1,77) = 3.70$, $p = .058$, indicating that students wrote their original essays more quickly over time. Students spent more time on their first few essays and gradually decreased. Similarly, students, on average, spent about 6 minutes revising their essays, with significant prompt differences over time. There was a significant linear effect, $F(1,75) = 24.96$, $p<.001$, a significant quadratic effect, $F(1,75) = 9.90$, $p = .002$, and a significant cubic effect, $F(1,75) = 6.28$, $p = .014$. Students spent more time revising their early essays and gradually decreased, with the exception of the Day 6 essay written on Uniformity. One interpretation is that high school students may have engaged with this topic more deeply because it was relevant to their daily lives (i.e., tension between fitting in versus being original) and prompted some students to make an extra effort. Overall, students appeared to take writing seriously, devoting the majority of their allotted time to the tasks.

Automated Scores

Essays were next examined with regard to the holistic scores assigned by the scoring algorithm. These analyses considered only students who provided a complete set of score data for all prompts and for both original and revised drafts (i.e., $n = 75$).

Students' essay scores for original and revised drafts were analyzed in a 2 (draft) x 8 (prompt) repeated-measures ANOVA (see Table 3). Both draft and prompt were within-subjects variables. For the prompt, there was a significant linear effect, $F(1,74) = 9.24$, $p = .003$. Over the course of training, students improved their overall essay scores, but the progression was not smooth. The Planning (Day 1) and Patience (Day 3) prompts appeared to be the most challenging for students, whereas performance seemed strongest for the Heroes (Day 4) and Fame (Day 8) prompts. More importantly, there was a significant main effect for draft, indicating that revised drafts were overall scored more highly than original drafts, $F(1,74) = 15.42$, $p<.001$. It is crucial to note, however, that these holistic improvements were rather small. Students were only given 10 minutes to implement their revisions – a constraint that was intended to reflect the tight time restrictions students might face on a standardized test. Thus, it is encouraging that students were able to improve their essays, even in minor increments, under such conditions.

There was a significant cubic interaction between the draft and prompt variables, indicating that essays written on particular prompts improved more than others, $F(1,74) = 5.85$, $p = .018$. Follow-up t-tests revealed that students made marginal or significant gains for the Winning (Day 2), Heroes (Day 4), Beliefs (Day 7), and Fame (Day 8) prompts (see Table 3).

TABLE 3
Mean essay drafting and revising durations and automated scores

	Duration in Minutes		**Score**			
Prompt	**Original**	**Revised**	**Original**	**Revised**	***t***	***p***
Day 1: Planning	23.4 (3.1)	7.8 (3.0)	2.36 (0.91)	2.32 (0.92)	−0.72	.471
Day 2: Winning	21.6 (4.6)	6.5 (3.3)	2.57 (0.95)	2.75 (0.81)	2.41	.019
Day 3: Patience	21.3 (4.9)	6.1 (2.9)	2.23 (0.98)	2.28 (0.98)	0.68	.496
Day 4: Heroes	21.8 (4.2)	5.7 (3.1)	2.96 (1.13)	3.25 (1.14)	3.48	.001
Day 5: Perfection	21.0 (5.3)	5.6 (3.2)	2.53 (1.18)	2.53 (1.20)	0.00	1.00
Day 6: Uniformity	21.1 (5.0)	6.0 (3.0)	2.62 (1.01)	2.71 (0.98)	1.18	.242
Day 7: Beliefs	21.0 (4.9)	5.5 (2.9)	2.52 (1.07)	2.65 (1.10)	1.92	.058
Day 8: Fame	20.6 (4.8)	5.5 (2.8)	2.72 (1.08)	2.92 (1.12)	3.51	.001

Automated Detection of Word-level Revisions

From previous analyses, it was clear that students did not respond to each prompt in the same manner. Some prompts inspired more effort and other prompts seemed more difficult. This is not surprising because the prompts, although they follow a similar style and format, necessarily deal with different themes and knowledge. Thus, in the following analyses of word-level revisions, we examined each prompt separately (see Table 4).

The overall trend observed for word-level revisions, as assessed by selected indices, was that students did not often implement many revisions. For example, when revising their Planning essay, students used fewer familiar words and first-person pronouns, while also incorporating more precise terms (i.e., decreased mean polysemy of words). For the Heroes essay, students used more concrete wording in their revisions but also shifted towards a less objective tone (i.e., increased second person pronouns). On the Beliefs essays, students improved their word choice by using more uncommon and unfamiliar words (i.e., decreased incidence of frequent and familiar words). Students also decreased the use of first-person pronouns in their writing, which often helps to establish a more objective tone. Thus, students may have sometimes used word-level revisions to express ideas more precisely.

Despite some work on the Planning and Beliefs essays, most essays demonstrated much less word-level revising. On average, essays written on the Winning,

TABLE 4
Summary of significant ($p<.05$) and marginal ($p\leq.07$) word-level essay revision patterns

Essay (Day)	Variable	Change	*F*	*p*	*d*
Day 1: Planning	Word familiarity	decrease	4.91	.030	−.087
	Word frequency	decrease	4.74	.032	−.067
	Polysemy	decrease	4.67	.034	−.115
Day 2: Winning	none				
Day 3: Patience	none				
Day 4: Heroes	Word concreteness	increase	5.49	.021	.078
	Second person pronouns	increase	4.30	.041	.058
Day 5: Perfection	none				
Day 6: Uniformity	Concreteness component	decrease	6.80	.011	−.104
Day 7: Beliefs	Word frequency	decrease	5.29	.024	−.063
	First person pronouns	decrease	5.23	.025	−.050
	Word familiarity	decrease	4.27	.042	−.047
Day 8: Fame	First person pronouns	increase	4.93	.029	.081

Patience, Perfection, Uniformity, and Fame prompts exhibited either a single significant revision or no changes. Thus, in contrast to previous research, we did not observe a dominant pattern of primarily word-level revisions. One possibility is that students in this study did not revise much at all. Despite making edits to their essays, perhaps the net result of these edits was minimal and not detectable. To further explore this possibility, we next consider document-level revisions, which tend to be less common for many student writers.

Automated Detection of Document-level Revisions

In contrast to revisions at the word level, significant document-level changes were observed across all writing prompts (Table 5). Two commonly observed revisions were increases in the overall number of words and number of paragraphs. When revising, students almost universally elaborated upon their existing ideas to add more details, examples, and other content, significantly increasing the length of their essays. In almost all cases, these additions were not merely a few details spread throughout existing paragraphs, but were communicated through new paragraphs or improved paragraph structure, such as revising a one-paragraph essay to use a clearer introduction-body-conclusion organization.

Other frequently-occurring revisions were detected via LSA measures comparing portions of the text. LSA assesses the semantic similarity of texts; an increased LSA score means that two texts (or portions of a text) have become more semantically related. For nearly every essay (Planning, Patience, Perfection, Uniformity, Beliefs, and Fame), we observed an increase in the LSA values comparing essay paragraphs to each other. Thus, in their revisions, students may have better tied their ideas together across sections of their essays. Rather than describing examples, anecdotes, or other ideas with little connection, perhaps these ideas were more explicitly cohesive.

Other indicators of improved essay cohesion were also observed (e.g., LSA given/new and referential cohesion). LSA given/new assesses how much information in one section of the text can be recovered or represented by preceding text. Thus, increases in this measure suggest that students were improving the flow of their writing such that new ideas more clearly built upon previous ideas rather than jumping from topic to topic. Such changes were exhibited in the Winning, Patience, Heroes, Perfection, Uniformity, and Fame essays. Similarly, the referential cohesion component measure assesses the degree to which similar ideas are referenced throughout a complete text. Essays written on the Heroes and Uniformity prompts demonstrated improvements in referential cohesion.

Some measures demonstrated a decrease in value (Table 5). For example, LSA scores for adjacent sentences (Beliefs essay) or all sentences (Heroes) both

TABLE 5
Summary of significant ($p<.05$) and marginal ($p\leq.07$) document-level essay revision patterns

Essay (Day)	Variable	Change	*F*	*p*	*d*
Day 1: Planning	Number of words	increase	35.50	<.001	0.34
	Number paragraphs	increase	15.81	<.001	0.31
	LSA paragraph-to-paragraph	increase	3.43	.068	0.17
Day 2: Winning	Number of words	increase	77.14	<.001	0.36
	Number of paragraphs	increase	13.24	<.001	0.14
	LSA given/new	increase	3.45	.067	0.07
Day 3: Patience	Number of words	increase	67.98	<.001	0.36
	Number of paragraphs	increase	16.82	<.001	0.24
	LSA given/new	increase	15.48	<.001	0.14
	LSA paragraph-to-paragraph	increase	4.75	.032	0.10
	Referential cohesion	increase	4.45	.038	0.09
Day 4: Heroes	Number of words	increase	82.09	<.001	0.38
	Number of paragraphs	increase	16.09	<.001	0.25
	LSA given/new	increase	4.59	.035	0.08
	LSA sentence-to-sentence	decrease	3.59	.061	−0.07
Day 5: Perfection	Number of words	increase	45.61	<.001	0.26
	LSA given/new	increase	10.38	.002	0.18
	Number of paragraphs	increase	7.52	.007	0.15
	LSA paragraph-to-paragraph	increase	7.27	.008	0.13
Day 6: Uniformity	Number of words	increase	71.08	<.001	0.44
	Number of paragraphs	increase	23.41	<.001	0.26
	LSA given/new	increase	19.39	<.001	0.14
	Narrativity component	increase	5.04	.027	0.06
	LSA paragraph-to-paragraph	increase	4.92	.029	0.14
	Referential cohesion	increase	4.35	.040	0.05
Day 7: Beliefs	Number of words	increase	69.88	<.001	0.38
	Number of paragraphs	increase	10.96	.001	0.38
	LSA adjacent sentences	decrease	4.08	.047	−0.07
	LSA paragraph-to-paragraph	increase	3.68	.058	0.07
Day 8: Fame	Number of words	increase	39.42	<.001	0.30
	LSA given/new	increase	7.42	.008	0.10
	LSA paragraph-to-paragraph	increase	4.39	.039	0.14

decreased. Importantly, both of these measures capture cohesion at a very local or superficial level, such as repeating the same word choices across many sentences. Although word repetition improves surface cohesion (i.e., ideas are clearly linked because the same ideas are reiterated over and over again), it is also indicative of a less-proficient writing style (McNamara, Crossley, & McCarthy, 2010). Thus, decreases in these measures represent essay improvements through the use of more varied ideas or expressions of those ideas.

In sum, automated assessments of document-level revisions suggest that students were revising their essays to communicate their ideas in a more *elaborated*, *organized*, or *cohesive* manner. Taken together, results from both word and

document-level revisions suggest that formative feedback from W-Pal helped to orient students toward high-level revising rather than more typical superficial edits. It is also worth noting that the effect sizes (i.e., Cohen's *d*) were generally small for most individual measures, which further highlights the incremental nature of revising. Students were not necessarily making sweeping changes during the revision period. Nonetheless, some of these revisions were detectable by our automated tools.

Example Revision

To provide a concrete example of revising that addresses high-level, elaborative issues, we consider an example essay written in response to the following Heroes prompt, which asked students to argue whether celebrities should be admired in the same manner as heroes:

> Having many admirers is one way to become a celebrity, but it is not the way to become a hero. Heroes are self made. Yet, in our daily lives, we see no difference between "celebrities" and "heroes." For this reason, we deprive ourselves of real role models. We should admire heroes, people who are famous because they are great, but not celebrities, people who simply seem great because they are famous. Should we admire heroes but not celebrities?

Original Draft

The original essay comprised three paragraphs and 350 words – it was of reasonable length and conformed to the most basic structure of essays: introduction, body, and conclusion. The essay began with an introduction that posed thought-provoking questions to engage the reader. For example, the student wrote, "Think of the person you most admire. Now, why do you admire them? Is it because other people admire them or that you feel like you should too? Or do you admire them for the things they do and the kind of person they are?" The introduction previewed key arguments and examples, such as the heroism of firefighters. Finally, the introduction concluded with a strong thesis that clearly stated the author's position, "Heroes should be admired because they improve the community around them and inspire altruism in others by being a valuable role model."

Despite a relatively strong introduction, the body and conclusion of the essay were under-developed. The single body paragraph related to the themes of community and role models and briefly provided additional details regarding firefighters and foster families. For example, the author wrote, "There are some heroes like firefighters that risk their lives in order to save others. They should be admired because of their bravery and caring for others. Other heroes like families

who open up their homes to take in foster children also make a difference by showing their compassion." These two examples might have been stronger if they had been developed separately within their own body paragraphs. The concluding paragraph was quite brief but generally restated the thesis and linked to broader themes of altruism and the public good. The final paragraph is excerpted below:

> Heroes should also be admired because they inspire other people and set a good example to be followed. When have a hero that has made a significant impact in our lives, it moves us to want to do something good too. We are influenced by others around us. If everyone admired heroes, we would have a more altruistic community. We would want to help others in spite of ourselves.

Feedback Received

The W-Pal feedback system correctly determined that this student likely did not need feedback on the introduction building strategies for this essay. Instead, the primary formative feedback offered (i.e., the Initial Topic) pertained to body building strategies and supporting ideas with additional factual evidence. The student also chose to request optional feedback on the Next Topic, which pertained to conclusion building strategies. This feedback encouraged the development of a clear conclusion, such as providing strategies for restating the thesis. Importantly, none of the feedback recommendations addressed factors such as spelling, grammar, punctuation, or similar low-level details. The formative feedback examples shown previously in Figure 1 were obtained from this student's individual feedback report.

Revised Draft

During the brief (10 minute) revision period, and in response to the given feedback, the student opted to focus her efforts on the final paragraph, with no changes made to the first two paragraphs. The last paragraph was transformed substantially into a new body paragraph that offered additional evidence in the form of a new example about a leukemia survivor and charity founder. Although problems remain – the example should have been developed with more detail – it represents a novel conceptual contribution to the essay. A new conclusion was also added that incorporated transitional wording (i.e., "overall") to signal the close of the essay. The new conclusion more precisely summarizes the original thesis regarding community impact and role models. Thus, the student appears to have taken the recommended thesis-restatement strategies to heart. The excerpt below showcases the revised paragraphs, with revised material indicated by italicized text.

Heroes should also be admired because they inspire other people and set a good example to be followed. *They put other people's needs above their own and do the right thing. By setting this example*, it *inspires other people* to want to *make a* good *contribution* too. *In recent news, there have been children who become heroes themselves. One child overcame leukemia, and now they started a fund to help other cancer patients who struggle with it. This can touch someone's life enough to make them go out and make a difference themselves because they see that a child can do it, so they can too.* If everyone admired heroes *rather*, we would have a more altruistic community. *Heroes serve as significant role models, instilling good values into us all.*

Overall, less attention should be paid to celebrities and more should be given to the heroes who make a difference in our lives. They deserve to be admired because of the impact they can have on a community and the valuable role they play in inspiring others.

Potential Feedback

The sample student's revisions exemplify the patterns observed in the computational linguistic analysis. Significant changes were made to add cohesive content (i.e., a substantive revision that affects the whole essay) but fewer changes were made to improve word choice or usage. In practice, writers often need to make meaningful revisions at all levels. In addition to feedback on strengths and weaknesses in the original essay alone, subsequent formative feedback might take these revision patterns into account. For instance, by using automated detection algorithms for word-level features, we could help this student implement more diverse, meaningful, or impactful wording. Instead of frequently repeating words (e.g., "admire" and "inspire") or relying upon relatively vague words (e.g., "good" and "people"), the student could be provided with strategies for identifying more compelling lexical alternatives.

DISCUSSION

For the development of the W-Pal intelligent tutoring system, a fundamental aim has been to strengthen students' writing proficiency via formative feedback that makes the knowledge, processes, and strategies of writing explicit and actionable. For example, such feedback may provide strategies for generating and elaborating ideas, developing compelling and objective arguments, or linking main ideas and themes to build deeper textual cohesion. Typically, such automated

feedback is offered in relation to data from a discrete essay. That is, an essay is written and submitted to the system, and resulting scores or feedback are bound by data (e.g., linguistic features) obtained from that specific draft. To expand the kinds of automated formative feedback that can be offered to students, we propose that it would be worthwhile to incorporate data related to transformations or revisions from draft to draft. Given that formative feedback may be most powerful when it addresses *processes* rather than only products (Shute, 2008), we hypothesize that this approach may ultimately improve the efficacy of automated writing instructional systems such as W-Pal.

A crucial first step in this process was to establish how and whether automated tools could detect meaningful revision patterns in students' essays. Could useful revision data be reliably extracted from students' writing using computational linguistic tools? In this paper, we demonstrated that automated detection of revising is both possible and informative. Importantly, we were able to detect meaningful revision patterns even under circumstances that might reasonably thwart substantive revising – writing under a tight time limit (10 minutes or less) with high school students. Using Coh-Metrix and related tools, we examined revisions at both the word-level and document-level in a corpus of almost 700 essays. Both types of revisions were observed, overall, but students were more likely to demonstrate document-level revisions related to elaboration and cohesion in response to W-Pal feedback. That is, students' tendency to only make lower-level edits (Bridwell, 1980; Crawford, Lloyd, & Knoth, 2008) seemed somewhat remedied through instruction and feedback from an intelligent tutoring system for writing.

These data obtained regarding students' revising practices can be used to inform new and potentially more effective automated feedback. Not only can feedback discuss the strengths, weaknesses, and opportunities within a given essay, feedback algorithms might now be developed that incorporate data on students' writing processes. For students who exhibit many word-level changes and few document-level changes from draft to draft, praise might first be offered for improving word choice and diversity. These writers might also be offered strategies to further bolster such revising to ensure that new words are not merely chosen because they "sound impressive" but because they more effectively convey one's ideas. In addition, feedback recommendations could instruct students about the importance of improving elaboration and organization, and offer targeted strategies for developing ideas (e.g., freewriting and outlining), building cohesion and flow across paragraphs (e.g., threading ideas throughout the text and removing vague pronouns), and better explaining examples and evidence. Importantly, this individualized feedback can be carefully aimed at students

whose revising behaviors explicitly show that they have neglected such concerns when preparing their latest draft. Consequently, there may also be motivational benefits for students who perceive that the automated feedback is truly specific and personalized to their *own* writing (see Grimes & Warschauer, 2010; Roscoe & McNamara, 2013).

Several studies have improved revising through strategy instruction (e.g., Butler & Britt, 2011; Midgette, Haria, & MacArthur, 2008, Proske, Narciss, & McNamara, 2012). The automated detection of revisions may enhance the ability to guide when and what instruction is offered to the writer. For instance, Midgette and colleagues (2008) instructed middle school students to adopt one of several revising goals: generally improve, elaborate, or elaborate and consider the audience. Students taught to elaborate with the audience in mind were better able to revise their essays to address differing perspectives, although holistic scores did not vary by condition. By tracking indicators of elaborative revisions (e.g., added examples) and audience-sensitive revisions (e.g., added cohesive cues), we might determine when students would benefit from adopting modified writing goals. Similarly, Butler and Britt (2010) examined undergraduates' revising as a function of training: no training, global revising (i.e., substantive revisions of sentences, paragraphs, or whole text), argument revising (i.e., more precise language and addressing counterarguments), or both global and argument training. Students who received any training demonstrated more substantive revising and improved argument quality, whereas students who received no training focused on superficial edits. By examining linguistic revising data, such as changes to improve precision (e.g., mean hypernymy) or changes at multiple textual levels, automated writing instruction tools might provide individualized training that targets writers' weak areas.

A further benefit of this approach echoes other strengths of automated writing evaluation related to time constraints (e.g., Dikli, 2006; Shermis & Burstein, 2013; Warschauer & Grimes, 2008). Specifically, classroom teachers often give feedback on students' writing and revising, but such grading and feedback require substantial time. Students may no longer be "in the moment" by the time the feedback is received. Students may have forgotten their intentions or goals during the revision process, and thus feedback on such revisions may lack impact or crucial context. Although writing feedback need not be immediate to be effective, is important that students can connect the feedback received to their personal thinking and writing processes (Shute, 2008; Nicol & Macfarlane-Dick, 2006). With the ability to add immediate or real-time feedback on the revising process to automated writing tools (using data extracted from students' own essays), we may further strengthen the affordances of these technologies. Research with other

intelligent tutoring tools has found that real-time, individualized feedback can improve students' learning processes and strategies, such as help-seeking strategies (Roll, Aleven, McLaren, & Koedinger, 2011) and peer tutoring strategies (Walker, Rummel, & Koedinger, 2011). Thus, when students receive feedback on their own revising activities while the cognitive processes and goals of these activities are still engaged, active, and malleable, it may help them to better appreciate the benefits (or lack of benefits) of their revisions.

One remaining question from this study pertains to prompt-based differences. Results suggested that certain prompts are perhaps easier for students to write about than others, and may promote different patterns of writing or revising. Unfortunately, it was beyond the scope of the current work to tease apart what features or content in the prompts specifically influenced students' writing styles. In future research, it may be powerful to combine linguistic data from essays with students' own self-reported perceptions of the prompts and their writing. Did students enjoy the topic or feel a personal connection to the essay, or did students find the prompt concepts unfamiliar or boring? Did students feel that they had produced their best work or a poor-quality text? By incorporating such perceptual, attitudinal, or other individual difference data, we may be able to better predict how and whether students will revise, and perhaps interpret how students' enacted revisions were linked to their personal writing goals.

Another limitation of the current research is that only a subset of possible linguistic indices were evaluated at the word and document levels. This was necessary to conduct a tractable analysis and provide a proof-of-concept. However, moving forward, there are few restrictions on incorporating more indices from multiple computational sources. Assessing essays using more measures should provide even more detailed portraits of students' revising patterns. Similarly, we focused here on word and document levels of data, but meaningful revisions can and do occur at intermediate sentence and paragraph levels. Indeed, in the student example essay that we reviewed, the student's revisions were confined to two paragraphs and involved sentence-level revisions. As additional layers of analysis and complexity are added to this methodology, we expect to gain further insight into how students revise in response to automated feedback.

A related concern relates to how particular linguistic variables are classified as "word-level" or "document-level" revision indicators. For example, we have situated pronoun revisions (i.e., changes in the use of first, second, and third-person pronouns) within the word level. This categorization was based on the assumption that when students revise their pronouns (e.g., changing "when *you* believe" to "when *he or she* believes"), they are most often making local edits to the specific wording of their text. However, one could argue that when such changes are

made throughout a text, it is indicative of substantive changes in overall tone, perspective, or objectivity (i.e., document-level). Thus, for certain variables, the extent to which given revision patterns are interpreted as superficial or substantive may depend on the distribution or systematicity of observed edits. In future work, more precise models for analyzing and interpreting revision data can be specified, and these models can ultimately be linked to changes in holistic text quality or writing proficiency.

ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, US Department of Education, through Grants R305A120707 and R305A090623 to Arizona State University. The opinions expressed are those of the author and do not represent views of the Institute or the US Department of Education. The authors would also like to thank Kaitlyn Long for her assistance.

HIGHLIGHTS

- automated writing evaluation uses linguistic data to assess student writing
- computational linguistic data can also be used to guide feedback on writing
- Writing Pal users improve their essays based on automated formative feedback
- extends current work by demonstrating automated detection of essay revising
- automated feedback can now be expanded to address student revising patterns

REFERENCES

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*. Available from http://www.jtla.org.

Bridwell, L. (1980). Revising strategies in twelfth grade students' transactional writing. *Research in the Teaching of English, 14*, 197–222.

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

Butler, J. A., & Britt, M. A. (2011). Investigating instruction for improving revision of argumentative essays. *Written Communication, 28,* 70–96.

Camara, W. J. (2003). *Scoring the essay on the SAT writing section.* New York: College Entrance Examination Board.

Crawford, L., Lloyd, S., & Knoth, K. (2008). Analysis of student revisions on a state writing test. *Assessment for Effective Intervention, 33,* 108–119.

Crossley, S. A., Roscoe, R. D., & McNamara, D. S. (2013). Using automatic scoring models to detect changes in student writing in an intelligent tutoring system. In P. M. McCarthy & G. M. Youngblood (Eds.), *Proceedings of the 26th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 208–213). Menlo Park, CA: The AAAI Press.

Crossley, S. A., Varner, L. K., Roscoe, R. D., & McNamara, D. S. (2013). Using automated cohesion indices as a measure of writing growth in intelligent tutoring systems and automated essay writing systems. In K. Yacef et al. (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED) Conference* (pp. 269–278). Heidelberg, Berlin: Springer.

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., & Bivens-Tatum, J. (2008). *Cognitive models of writing: Writing proficiency as a complex integrated skill* (Research Report No. RR-08-55). Princeton, NJ: Educational Testing Service.

Dikli, S. (2006). An overview of automated essay scoring of essays. *Journal of Technology, Learning, and Assessment*, *5,* Available from http://www.jtla.org.

Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research, 57*, 481–506.

Flower, L., & Hayes, J. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*, 365–387.

Grimes, D. & Warschauer, M. (2010). Usability in a fallible tool: a multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment, 8*, Retrieved from http://www.jtla.org.

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., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). ***Handbook of Latent Semantic Analysis.*** **Mahwah, NJ: Erlbaum.**

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

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*, 57–86.

McNamara, D. S., Crossley, S. A., & Roscoe, R. D. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods, 45*, 499–515.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. Assessing Writing, *23*, 35–59.

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 188–205). **Hershey, PA: IGI Global.**

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* **Cambridge: Cambridge University Press.**

McNamara, D. S., Raine, R., Roscoe, R. D., Crossley, S., Jackson, G. T., Dai, J., Cai, Z., Renner, A., Brandon, R., Weston, J., Dempsey, K., Carney, D., Sullivan, S., Kim, L., Rus, V., Floyd, R., McCarthy, P. M., & Graesser, A. C. (2012). The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 298–311). **Hershey, P.A.: IGI Global.**

Midgette, E., Haria, P., & MacArthur, C. (2008). The effects of content and audience awareness goals for revision on the persuasive essays of fifth- and eighth-grade students. *Reading and Writing, 21,* 131–151.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Students in Higher Education*, *31*, 199–218.

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction, 21*, 267–280.

Roscoe, R. D., Brandon, R. D., Snow, E. L., & McNamara, D. S. (2013). Game-based writing strategy practice with the Writing Pal. In K. Pytash & R. Ferdig (Eds.), *Exploring technology for writing and writing instruction* (pp. 1–20) Hershey, PA: IGI Global.

Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology, 105*, 1010–1025.

Roscoe, R. D., Snow, E. L., & McNamara, D. S. (2013). Feedback and revising in an intelligent tutoring system for writing strategies. In K. Yacef et al. (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED)* (pp. 259–268). Heidelberg, Berlin: Springer.

Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition 34*, 39–59.

Shermis, M. D., & Burstein, J. C. (Eds.). (2013). *Handbook of automated essay evaluation: current applications and new directions*. New York, NY: Routledge.

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

Sommers, N. (1980). Revision strategies of student writers and experienced adult writers. *College Composition and Communication, 31*, 378–388.

Walker, E., Rummel, N., & Koedinger, K. R. (2011). Designing automated adaptive support to improve student helping behaviors in a peer tutoring activity. *International Journal of Computer-supported Collaborative Learning*, *6*, 279–306.

Warschauer, M. & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, *3*, 22–36.