Feedback and Revising in an Intelligent Tutoring System for Writing Strategies

Rod D. Roscoe[1], Erica L. Snow[1,2], and Danielle S. McNamara[1,2]

[1] Learning Sciences Institute
[2] Department of Psychology, Arizona State University
{rod.roscoe,erica.l.snow,danielle.mcnamara}@asu.edu

Abstract. This study investigates students' essay revising in the context of an intelligent tutoring system called *Writing Pal* (W-Pal), which combines strategy instruction, game-based practice, essay writing practice, and automated formative feedback. We examine how high school students use W-Pal feedback to revise essays in two different contexts: a typical approach that emphasizes intensive writing practice, and an alternative approach that offers less writing practice with more direct strategy instruction. Results indicate that students who wrote fewer essays, but received W-Pal strategy instruction, were more likely to make substantive revisions that implemented specific recommendations conveyed by the automated feedback. Additional analyses consider the role of motivation and perceived learning on students' revising behaviors.

Keywords: intelligent tutoring systems, writing instruction, writing strategies, automated feedback, natural language processing, motivation.

1 Introduction

Writing is a complex process comprising planning, drafting, and revising phases [1-2]. Planning refers to the generation and organization of ideas prior to writing and drafting translates writers' initial ideas into a coherent text that communicates main ideas. Central to the current work, revising entails the refinement of a text to better achieve writers' goals. Skilled writers engage in more substantive revising that addresses deeper organization, meaning, and rhetorical strength (e.g., elaborating and restructuring arguments), which is more likely to improve overall essay quality [3]. However, many students tend to ignore revising or make only unproductive, superficial edits to address spelling, grammar, and mechanical issues [3-6].

Writing Pal (W-Pal) is an intelligent tutoring system developed to improve students' writing and revising [7-8]. Via animated lessons and educational games, W-Pal offers explicit strategy instruction and practice for planning, drafting, and revising. Importantly, students can also author essays and receive automated formative feedback informed by natural language processing (NLP) algorithms [9]. In this study, we investigate students' use of such feedback to revise their essays. Specifically, we consider whether and how students can use automated feedback to guide substantive revisions, and how revising may be influenced by explicit strategy instruction.

1.1   Revising and Computers

Research on revising indicates that many students rely on superficial edits rather than substantive revisions [3-6]. For example, Bridwell [4] analyzed Grade 12 students' essay revisions at seven grain sizes: surface, words, phrases, clauses, sentences, multiple sentences, and text level. All students revised, but most revisions occurred at the word (31.2%) or surface level (24.8%). Students revised primarily by improving word choice and by correcting mechanical errors. Similarly, Crawford et al. [5] examined the revisions of Grade 5 and Grade 8 students. These elementary and middle school students' revisions also focused on the word (~40%), level (~25%), or punctuation level (~20%), although these edits did lead to moderate increases in essay quality.

Efforts to improve students' revising processes have focused on strategy instruction [3, 10-11] and computer-based scaffolds [12-13]. For example, Midgette et al. [11] provided Grade 5 and Grade 8 students with one of three revising goals: generally improve, elaborate the content, or elaborate the content and consider the audience. Students given an audience goal were better able to revise their essays to address alternative perspectives (i.e., substantive revisions), although essay quality did not differ across conditions. Similarly, Butler and Britt [10] analyzed the revisions of undergraduates given no training, a global revision tutorial (i.e., substantive revisions of sentences, paragraphs, or whole text), an argument revision tutorial (i.e., precise language and addressing counterarguments), or both tutorials. Students who received either tutorial engaged in more substantive revising and improved overall argument quality, whereas students who received no training focused on less-productive superficial edits. Thus, strategy instruction appears to facilitate substantive essay revising.

Other research has explored the benefits of automated writing evaluation (AWE) systems that combine automated scoring with error feedback [12-14]. Such systems seek to improve students' writing and revising by enabling substantially more writing practice than is often feasible given classroom time constraints [13]. In practice, research on AWE has focused on scoring accuracy. Human and computer-assigned scores correlate around .80 to .85, and many systems report 40-60% perfect agreement between human and computer scores, and 90-100% adjacent agreement (i.e., scores within 1 point) [12, 15]. However, accurate scoring does not guarantee that students are able to implement the feedback. For example, *Criterion* [16] utilizes NLP and statistical modeling to automatically score essays and generate feedback related to errors of organization, development, grammar, usage, mechanics, and style. Attali [17] investigated *Criterion* with thousands of Grade 6 through Grade 12 students – over 33,000 essays were submitted to the system. Most of these essays (71%) were not revised. However, analyses showed that students who did revise implemented superficial edits along with occasional substantive revisions to discourse elements.

As computer-based supports for writing gain educational and commercial prominence, it is crucial to explore whether and how students can use automated feedback to revise their essays. Moreover, it is important to consider how explicit strategy instruction and AWE can be synthesized to support revising. To address these questions, we examine essay revising in the context of the W-Pal tutoring system.

1.2   Writing Pal

W-Pal offers writing strategies via eight writing modules comprising instructional videos, narrated by pedagogical agents, and educational practice games (Table 1). The videos provide background information about key writing tasks (e.g., writing a thesis) and decompose the goals and operations for each strategy. Multiple strategies are often organized by acronymic mnemonic devices, which can facilitate students' recall and use of the strategies [18]. Completing the lessons unlocks games that allow students to practice specific strategies. In *identification* games, students examine short texts and essay excerpts to identify strategy applications or exemplars. For example, in *Fix-It*, players attempt to identify problems exhibited in introduction, body, or conclusion paragraphs. In *generative* games, students author short texts while applying one or more strategies. For example, in *Speech Writer*, players help a friend on the debate team by reviewing a "speech" for key problems and then revising that speech.

Table 1. Writing Pal (W-Pal) Writing Strategy Modules, Lesson Videos, and Practice Games

Module	Strategy Lessons	Practice Games
Prologue	*Meet the Student* *Practice Makes Perfect*	
Freewriting	*Figure Out the Prompt* *Ask and Answer Questions* *Support with Evidence* *Think about the Other Side*	*Freewrite Flash*
Planning	*Positions, Arguments, and Evidence* *Outlines* *Flowcharts*	*Planning Passage* *Mastermind Outline*
Introduction Building	*Thesis Statements* *Argument Previews* *Grab the Reader's Attention*	*Essay Launcher* *Dungeon Escape* *Fix It*
Body Building	*Topic Sentences* *Evidence Sentences* *Strengthening Your Evidence*	*RoBoCo* *Fix It*
Conclusion Building	*Summarize the Essay* *Close the Essay* *Hold the Reader's Attention*	*Lockdown* *Dungeon Escape* *Fix It*
Paraphrasing	*Synonym Strategy* *Structure Strategy* *Condensing Strategy* *Splitting Strategy*	*Adventurer's Loot* *Map Conquest*
Cohesion Building	*Signpost Strategy* *Threading* *Connectives Strategy*	*Undefined & Mined* *CON-Artist*
Revising	*Add More* *Removing Irrelevant Details* *Moving Essay Sections* *Substituting Ideas*	*Speech Writer*

Similar to AWE systems, W-Pal also allows students to write and revise prompt-based essays like those on standardized exams. Essays are automatically scored via NLP algorithms developed using Coh-Metrix and related tools [9], which provide a key source of the artificial intelligence of the system. Within technologies that accept natural language as input, students' responses are open-ended and potentially ambiguous. When a user enters natural language into a system and expects useful and intelligent responses, NLP is necessary to interpret that input. In service to these goals, W-Pal utilizes Coh-Metrix to analyze text on multiple dimensions, including co-referential cohesion, causal cohesion, density of connectives, lexical diversity, temporal cohesion, spatial cohesion, and LSA. Coh-Metrix also calculates syntactic complexity and offers psycholinguistic data about words (parts-of-speech, frequency, concreteness, imagability, meaningfulness, familiarity, polysemy, and hypernymy). A variety of methods, including regression, discriminant function analysis, and machine learning, are used to combine indices in models that assign scores (or qualitative thresholds) to essays as a whole or essay sections (e.g., a conclusion paragraph).

In W-Pal, submitted essays receive a holistic rating from *Poor* to *Great* (6-point scale). Essays then receive formative feedback on specific writing goals and strategies, implemented through a series of algorithmic thresholds assessing *Legitimacy, Length*, *Relevance*, *Structure*, *Introduction*, *Body*, *Conclusion*, or *Revising*. Unlike most AWE systems, W-Pal provides no feedback on low-level errors and provides less feedback overall to avoid overwhelming users [14]. W-Pal automatically gives one feedback message on one *Initial Topic* (i.e., the *first* problem detected in the series of checks). Subsequently, students can voluntarily request more feedback on that topic or on one additional *Next Topic* (i.e., the *next* problem detected). Up to ten total feedback messages, five per topic, can be requested by the students. Below is an example of a complete feedback message on the topic of conclusion building:

Skilled writers attempt to hold the reader's attention throughout each segment of the essay. One way to ensure your essay conclusion is interesting to your reader is to use an attention-holding technique.

- These techniques help your reader connect to the essay on a personal level.
- A simple technique is to use personal stories that have not been previously discussed in the essay.
- Consider this prompt: "Is it always better to tell the truth?" A personal anecdote might discuss how, after having hurt your mom's feelings by telling a lie, you learned a lesson about honesty.

In sum, W-Pal strives to integrate strategy instruction and essay-based practice with automated feedback. We hypothesize that strategy instruction will facilitate revising [10-11] by providing students with concrete methods of implementing the automated feedback, and perhaps by influencing their perceived ability to do so [19]. Thus, in this study, we consider 1) whether and how students can use automated feedback to inform substantive essay revisions, and 2) how revising occurs in two contexts: a typical AWE approach that emphasizes intensive writing practice (i.e., writing many essays with automated feedback) and an alternative approach that offers significantly less writing practice (i.e., fewer essays) but with more direct strategy instruction. Additionally, we explore relationships between students' use of feedback to revise and their self-reported motivation and perceptions of the system.

2   Method

2.1   Participants

High school students (n = 65) 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 16, with 70.8% females. Ethnically, 6.2% of students identified as African-American, 15.4% as Asian, 24.6% as Caucasian, and 44.6% as Hispanic. Average grade level was 10.2 with 35.4% of students reporting a GPA of 3.0 or below. Most students self-identified as native English speakers (n = 38) although many self-identified as English Language Learners (ELL, n = 27). An analysis of prior writing ability found no difference between native speakers and ELLs, $t(62) = 1.05$, $p = .30$.

2.2   Procedures

Students in the *W-Pal condition* began each session by writing and revising *one* SAT-style persuasive essay and then completing one instructional module (i.e., total of 8 practice essays on different topics). Students were allotted 25 minutes to draft their essay and 10 minutes to revise after receiving feedback. Subsequently, they studied the strategy module of the day and played the educational games. In the *Essay condition* (n = 32), students wrote and revised *two* essays per session (i.e., 16 practice essays), but did not complete any lessons or games. Sessions lasted about 1.5 hours for both conditions with equivalent time on task.

2.3   Data and Coding

Corpus. Students wrote and revised a combined total of 770 essays. Original and revised drafts were contrasted using the Compare Documents tool in a popular word processing program, thus highlighting the additions, deletions, and alterations students made when revising. The automated essay scores assigned to original and revised drafts were logged along with the duration (i.e., time spent writing), number of feedback messages requested, and topics of feedback given.

Revisions. Students' edits were coded in three ways. First, we coded whether students attempted to revise by making *any* edits. Second, we examined whether students attempted substantive revisions to address the Initial Topic of feedback. Students' edits were coded based on whether they implemented any valid strategy to address the specified feedback topic. For example, if a student received feedback related to essay introductions, the essay would be coded as *revised* if an introductory paragraph was added, or if a relevant introductory component was added (e.g., a preview of arguments) or meaningfully modified (e.g., elaborating the thesis statement). To establish coding reliability, the second author and an undergraduate assistant independently coded 120 essays. Reliability of Initial Topic coding was $\kappa = .84$. Finally, the same coding was applied to revisions based on the Next Topic of feedback ($\kappa = .81$).

Daily Surveys. Students completed a motivation survey at the start of each session. Using a 6-point scale, students rated their *enjoyment of the most recent session*, *motivation to participate*, *desire to perform well*, *desire to compete with others*, *perceived learning of writing strategies*, and *perceived improvements in writing quality*. Higher ratings indicated more positive perceptions (e.g., higher enjoyment, greater perceived learning, etc.). These data allow us to consider whether students' motivations or perceptions of W-Pal might have influenced their willingness to revise their essays [19].

3   Results

3.1   All Essays

We first examined writing times, scores, feedback patterns, and revising for the entire corpus of 770 essays. These data are summarized in Table 2.

Table 2. Writing duration, scores, feedback, and revising for all essays

Variable	Mean or Percentage	SD
Duration (minutes)		
Original	21.2	4.7
Revised	5.7	3.1
Score		
Original	2.6	1.0
Revised	2.7	1.0
Feedback Requested		
Total Received	3.4	3.0
1 message[a]	48.5%	
2-5 messages[a]	34.4%	
6+ messages[a]	19.0%	
Revising		
Total Edits	12.0	10.8
Any Revision[a]	97.3%	
Initial Topic Revision[a]	44.1%	
Next Topic Revision[a]	53.8%	

Note. [a]These values indicate a percentage of all essays.

Duration and Scores. On average, students spent 21 minutes composing their original drafts and 6 minutes revising (Table 2). The average score for original drafts was 2.6, which increased very slightly but significantly to 2.7 after revising, $t(769) = 4.21$, $p < .001$, $d = .08$. This result suggests that students essays improved incrementally (i.e., in relation to specific details or features) rather than holistically.

Feedback. On average, students received 3 to 4 feedback messages per essay (Table 2). Because students received one message by default, these data indicate that many

students actively requested 2 to 3 additional messages. Six essays did not receive feedback due to system error. The most common Initial Topic categories were Body Building (53.5% of essays), Revising (13.1%), Length (10.6%), and Conclusion Building (7.1%). Students requested Next Topic feedback for 34.0% of their essays. Of the 262 essays that received Next Topic feedback, the most common categories were Revising (17.7%), Introduction Building (7.1%), and Conclusion Building (6.8%). One implication is that students rarely had serious problems with basic essay features such as structure. Rather, students needed help with specific sections of their essays, such as how to introduce, develop, and summarize their arguments.

Revising. Over 97% of essays exhibited some attempt to revise and students made an average of 12.0 edits per essay (Table 2). However, a smaller percentage of essays displayed *substantive revisions* in response to received Initial Topic (44.1%) or Next Topic feedback (53.8%). Overall, students rarely ignored the opportunity to revise, but implemented substantive strategy feedback from W-Pal about half of the time.

3.2   Effects of Instruction and Practice Context

Although all students received feedback, the nature of instruction and practice differed experimentally. The W-Pal condition received strategy lessons, educational games, and wrote eight practice essays with automated feedback. The Essay condition engaged in twice as much writing practice with feedback, but did not complete the lessons or games. In the following analyses, we consider whether revising patterns differed in these two contexts. Because each student composed multiple essays, data for each student were aggregated. This aggregation obscured some of the variance within students and reduced statistical power, but was necessary to use students as the unit of analysis and meet assumptions of independent observations.

Table 3. Comparison of writing duration, scores, feedback, and revising across conditions

Variable	Condition		$F(1,63)$	p
	W-Pal	Essay		
Duration (minutes)				
Original	22.1 (2.9)	20.7 (3.8)	2.63	.11
Revised	6.0 (2.3)	5.5 (2.0)	< 1.00	.35
Score				
Original	2.7 (0.7)	2.5 (0.6)		
Revised	2.8 (0.8)	2.6 (0.6)		
Feedback Requests	3.7 (2.7)	3.2 (2.3)	< 1.00	.44
Revising				
Total Edits	11.4 (8.5)	12.4 (7.1)	< 1.00	.62
Any Revision[a]	98.1 (5.5)	96.8 (4.5)	1.03	.32
Initial Topic Revision[a]	53.7 (30.4)	39.2 (18.8)	5.32	.02
Next Topic Revision[a]	56.0 (40.2)	43.1 (33.4)	1.44	.24

Note. [a]These values are average percentages. They indicate what percentage of students essays were revised in the indicated manner, on average.

Duration and Scores. On average, W-Pal students spent 22 minutes composing their original drafts compared to 21 minutes spent by Essay students. Similarly, W-Pal students spent about 6 minutes revising compared to 5.5 minutes spent by Essay students. Neither difference was statistically significant (Table 3).

A 2 x 2 repeated-measures, mixed-factor ANOVA was conducted to compare original and revised drafts scores (within) by condition (between). A main effect of revision indicated that scores increased very slightly after being revised, $F(1,63) = 13.26$, $p = .001$, $d = .12$. However, there was no effect of condition, $F(1,63) < 1.00$, and no interaction, $F(1,63) < 1.00$. Although essay quality slightly improved as a result of revising, neither condition improved more than the other (Table 3).

Feedback. The conditions did not differ significantly in feedback received. On average, W-Pal students received 3.7 messages and Essay students received 3.2 messages.

Revising. W-Pal and Essay groups made a similar number of edits. Likewise, W-Pal students revised their essays 98% of the time and Essay students revised their essays 97% of the time. For substantive revisions in response to received feedback, W-Pal condition students showed a clear advantage. In response to Initial Topic feedback, W-Pal students made substantive revisions 54% of the time whereas Essay students made substantive revisions only 39% of the time, $F(1,63) = 5.32$, $p = .024$, $d = .57$. In response to Next Topic feedback, W-Pal students made substantive revisions 56% of the time, whereas Essay students made substantive revisions 43% of the time. Although not significant, this followed the same trend as Initial Topic feedback ($d = .35$). The percentage of essays revised in response to Initial Topic ($r = .30$, $p = .015$) or Next Topic feedback ($r = .42$, $p = .003$) was correlated with revised essay scores.

In sum, the groups did not differ in writing time or overall revising, but students who received both explicit strategy instruction and essay-based practice seemed more likely or able to implement automated writing feedback than students who only engaged in intensive essay-based practice.

Table 4. Correlations between motivational ratings and revisions

	Revisions		
Ratings	Any	Initial Topic	Next Topic
Enjoyment of Recent Session	.18	.32[b]	.12
Motivation to Participate	.08	.19	.01
Desire to Perform Well	.06	.23	.05
Competitiveness	-.04	.10	-.07
Perceived Strategy Learning	.30[b]	.31[b]	.16
Perceived Writing Improvement	.34[a]	.25[b]	.10

Note [a] $p \leq .01$. [b] $p \leq .05$.

3.3   Role of Motivation

In further analyses, we considered how students' motivations may have influenced their revising. For each survey item, ratings were averaged across sessions to provide

an aggregate rating. Correlations were computed between ratings and students' mean percentage of implementing any revisions, substantive Initial Topic revisions, and substantive Next Topic revisions (Table 4). Due to a logging error, the data for one student in the Essay condition could not be used, reducing the sample size ($n = 64$).

In general, students who perceived that their writing strategies and essay quality were improving seemed more likely to make revisions. Substantive Initial Topic revisions were also moderately correlated with perceived learning and improvement, along with enjoyment of the training sessions. None of the ratings were correlated with substantive Next Topic revisions. Thus, students' perceptions seemed not to affect whether they implemented recommendations beyond the first topic.

4   Discussion

Computer-based writing instruction typically strives to increase the number of essays students write and revise [11]. In this study, we examined how and whether students can revise essays based on automated feedback and how strategy instruction might bolster revising. Results suggest that students can utilize automated formative feedback, and the combination of strategy instruction, educational games, and essay-based practice was more supportive of substantive revising than simply writing and revising many essays. Students in both groups interacted with the same W-Pal writing and feedback tools, and students were able to make small, incremental improvements in essay quality. Thus, the automated feedback provided by W-Pal, guided by natural language algorithms, was moderately helpful to high school students. However, users of the full W-Pal were more willing or able implement substantive revisions. Our interpretation is that strategy instruction and game-based practice helped students to better understand the feedback and how to respond. That is, knowledge of specific strategies helped students understand how to act upon the feedback recommendations.

Importantly, students who perceived that they were learning and improving were also somewhat more likely to revise and make substantive revisions. Strategy instruction perhaps helped students feel more capable in their ability to revise. Students may have been more willing to revise substantively because they felt more equipped to do so. Future research will need to explore how computer-based writing instruction may further encourage students' positive attitudes toward writing and revising.

References

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

2. Flower, L., Hayes, J.: A cognitive process theory of writing. College Composition and Communication 32, 365–387 (1981)
3. Fitzgerald, J.: Research on revision in writing. Review of Educational Research 57, 481–506 (1987)
4. Bridwell, L.: Revising strategies in twelfth grade students' transactional writing. Research in the Teaching of English 14, 197–222 (1980)
5. Crawford, L., Lloyd, S., Knoth, K.: Analysis of student revisions on a state writing test. Assessment for Effective Intervention 33, 108–119 (2008)
6. Sommers, N.: Revision strategies of student writers and experienced adult writers. College Composition and Communication 31, 378–388 (1980)
7. McNamara, D., Raine, R., Roscoe, R., Crossley, S., Jackson, G., Dai, J., Cai, Z., Renner, A., Brandon, R., Weston, J., Dempsey, K., Carney, D., Sullivan, S., Kim, L., Rus, V., Floyd, R., McCarthy, P., Graesser, A.: The Writing Pal: natural language algorithms to support intelligent tutoring of writing strategies. In: McCarthy, P.M., Boonthum-Denecke, C. (eds.) Applied Natural Language Processing and Content Analysis: Identification, Investigation, and Resolution, pp. 298–311. IGI Global, Hershey (2012)
8. Roscoe, R., Varner, L., Weston, J., Crossley, S., McNamara, D.: The Writing Pal intelligent tutoring system: usability testing and development. Computers and Composition (in press)
9. McNamara, D., Crossley, S., Roscoe, R.: Natural language processing in an intelligent writing strategy tutoring system. Behavior Research Methods (2012), doi:10.3758/s13428-012-0258-1
10. Butler, J., Britt, M.: Investigating instruction for improving revision of argumentative essays. Written Communication 28, 70–96 (2011)
11. Midgette, E., Haria, P., MacArthur, C.: 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 (2008)
12. Dikli, S.: An overview of automated scoring of essays. Journal of Technology, Learning, and Assessment 5 (2006), http://www.jtla.org
13. Shermis, M., Burstein, J.: Automated essay scoring: a cross-disciplinary perspective. Erlbaum, Mahwah (2003)
14. Grimes, D., Warschauer, M.: Utility in a fallible tool: a multi-site case student of automated writing evaluation. Journal of Technology, Learning, and Assessment 8 (2010), http://www.jtla.org
15. Warschauer, M., Ware, P.: Automated writing evaluation: defining the classroom research agenda. Language Teaching Research 10, 1–24 (2006)
16. Burstein, J., Chodorow, M., Leacock, C.: Automated essay evaluation: the Criterion online writing system. AI Magazine 25, 27–36 (2004)
17. Attali, Y.: Exploring the feedback and revision features of Criterion. Paper Presented at the National Council on Measurement in Education, San Diego (April 2004)
18. de la Paz, S., Graham, S.: Explicitly teaching strategies, skills, and knowledge: writing instruction in middle school classrooms. Journal of Educational Psychology 94, 687–698 (2002)
19. Pajares, F., Johnson, M., Usher, E.: Sources of writing self-efficacy beliefs of elementary, middle, and high school students. Research in the Teaching of English 42, 104–120 (2007)