Article

Journal of Educational Computing Research
2022, Vol. 60(3) 696–721
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/07356331211045304
journals.sagepub.com/home/jec

Strategy Uptake in Writing Pal: Adaptive Feedback and Instruction

Reese Butterfuss[1], Rod D. Roscoe[2], Laura K. Allen[3], Kathryn S. McCarthy[4], and Danielle S. McNamara[1]

Abstract

The present study examined the extent to which adaptive feedback and just-in-time writing strategy instruction improved the quality of high school students' persuasive essays in the context of the Writing Pal (W-Pal). W-Pal is a technology-based writing tool that integrates automated writing evaluation into an intelligent tutoring system. Students wrote a pretest essay, engaged with W-Pal's adaptive instruction over the course of four training sessions, and then completed a posttest essay. For each training session, W-Pal differentiated strategy instruction for each student based on specific weaknesses in the initial training essays prior to providing the opportunity to revise. The results indicated that essay quality improved overall from pretest to posttest with respect to holistic quality, as well as several specific dimensions of essay quality, particularly for students with lower literacy skills. Moreover, students' scores on some of the training essays improved from the initial to revised version on the dimensions of essay quality that were targeted by instruction, whereas scores did not improve on the dimensions that were not targeted by instruction. Overall, the

[1] Department of Psychology, Arizona State University, Tempe, AZ, USA
[2] Human Systems Engineering, Arizona State University - Polytechnic, Mesa, AZ, USA
[3] Department of Psychology, University of New Hampshire, Durham, NH, USA
[4] Department of Learning Sciences, Georgia State University, Atlanta, GA, USA

Corresponding Author:
Reese Butterfuss, Department of Psychology, Arizona State University, 950 S. McAllister Ave, Tempe, AZ 85287, United States.
Email: rmbutterfuss@gmail.com

results suggest that W-Pal's adaptive strategy instruction can improve the quality of students' essays overall, as well as more specific dimensions of essay quality.

Keywords

writing strategies, adaptive instruction, intelligent tutoring systems, automated writing evaluation

Effectively conveying meaning in written texts is critical for academic and professional success. Writing performance is among the best predictors of academic achievement (Graham, et al., 2020; Graham & Hall, 2016). However, students struggle with writing as a result of underdeveloped knowledge and skills (Miller et al., 2018; National Commission on Writing, 2004). In general, students do not receive sufficient writing instruction (Graham et al., 2014), in part because many instructors feel underprepared to teach writing (Kiuhara et al., 2009) and are often limited in their capacity to provide opportunities for writing practice along with targeted, individualized feedback. In response, researchers have developed technology-based writing tools. One such tool is the Writing Pal (W-Pal; Roscoe & McNamara, 2013), which provides writing strategy instruction, extended practice, and immediate formative feedback. W-Pal has demonstrated effectiveness for improving high school and college students' strategy knowledge and writing performance (e.g., Crossley et al., 2016; Roscoe et al., 2013, 2014, 2015; Roscoe & McNamara, 2013).

Despite the promise of W-Pal to improve students' writing performance, there remain opportunities to further develop its capabilities. Specifically, previous iterations of the system were not adaptive to students' performance, as there were no mechanisms for recommending instructional modules or sequences of modules based on the quality of students' essays. In this study, we implement and examine an adaptive instructional cycle that uses automated evaluation of students' essays to direct them to an instructional strategy module targeting specific weaknesses (e.g., introduction quality, body quality, or lack of elaboration) in those essays. After studying these modules, students have the opportunity to revise. Rather than engaging with *all* instructional modules within W-Pal, this adaptive approach individualizes instruction based on students' needs identified from their writing. The goal of the adaptive instruction, particularly when coupled with feedback and practice opportunities, is to address students' weaknesses and improve the quality of their writing overall. In the current study, we examined the extent to which W-Pal's new adaptive, just-in-time instruction improved the quality of high school students' persuasive essays. We were particularly interested in the extent to which W-Pal improved essay quality for students with lower literacy skills (i.e., reading comprehension), as these students are most likely to face writing difficulty. We also examined the extent to which W-Pal fostered improvement on the specific weaknesses identified in students' writing for which they received individualized instruction.

The Challenge of Developing Writing Proficiency

A core challenge of writing instruction lies in the multidimensional nature of writing, which requires students to coordinate several cognitive skills and knowledge sources through multiple demanding processes including setting goals, solving problems, and strategically managing cognitive resources (Allen et al., 2016; Flower & Hayes, 1981; Hayes, 1996). Every student brings a unique profile of skills to a given writing task, and these variables interact to influence students' writing processes and, in turn, the strategies and procedures they draw upon.

When students can successfully leverage these skills over the course of prewriting, drafting, and revising, then the end product is more likely to be a relatively high-quality essay (Flower & Hayes, 1980; Kellogg & Whiteford, 2009).

Successful essays and successful writers do not necessarily share the same linguistic attributes or characteristics as there are many ways to write a good essay and multiple profiles of successful writers (Crossley et al., 2014). Indeed, the quality of an essay cannot be defined by a singular set of pre-defined linguistic properties. Critical dimensions of writing, such as cohesion, structure, lexical sophistication, and content are all important linguistic elements of quality, but students combine them in different ways to produce successful writing (Crossley et al., 2014). A critical implication of this variability is that students can succeed or fail at writing tasks in various ways. Thus, students may benefit from instruction and feedback that target their specific needs (Dempsey et al., 2009; Ericsson et al., 1993). Students' needs may also vary across essays, as students may show patterns of strengths and weaknesses that differ across various essay prompts (Allen et al., 2016, 2019; Crossley, 2020). These strengths and weaknesses can also change over the course of writing practice, feedback, and instruction (Kieft et al., 2007; Torrance et al., 2000). Consequently, a one-size-fits-all approach to writing instruction and support may be less effective than an approach that individualizes feedback and instruction for each student at the level of the individual essay, as well as across essays.

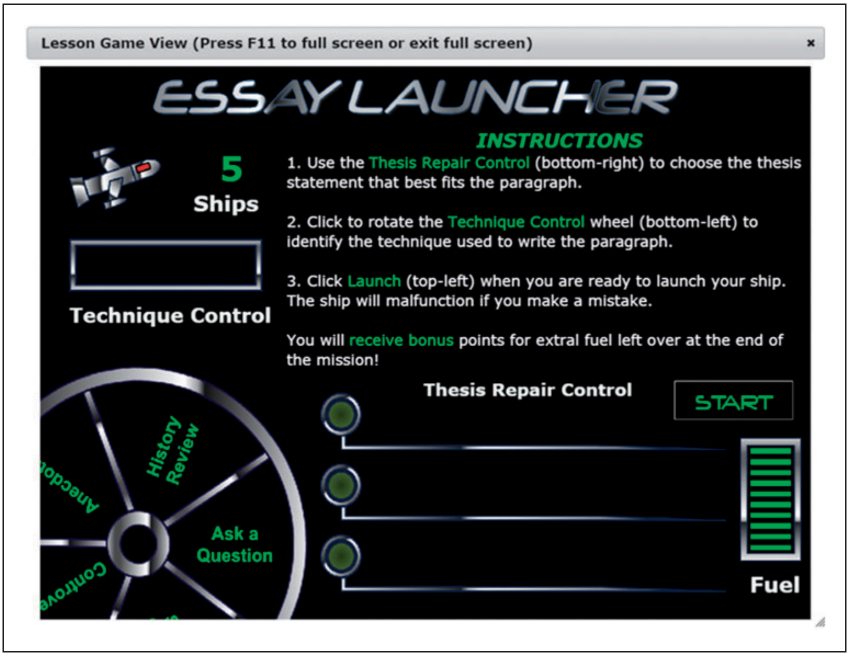
A substantial challenge for providing such feedback and instruction is teachers' ability to provide students with high-quality individualized feedback, instruction, and opportunities for extended practice, all of which are necessary for skill building in general (Ericsson et al., 1993), as well as for development of writing skills (Graham & Perin, 2007; Johnstone et al., 2002; Kellogg & Raulerson, 2007). Extended practice is much easier to develop and deploy in well-structured domains (e.g., times tables and vocabulary building) but far more complex and time consuming for ill-structured and open-ended tasks such as writing essays. Fortunately, researchers have developed technology-based writing tools that can efficiently provide students with instruction, practice opportunities, and feedback to improve writing performance.

Writing Pal: Combining Automated Writing Evaluation and Intelligent Tutoring

Writing Pal (W-Pal) was designed to provide writing strategy instruction, feedback, and extended practice opportunities for developing adolescent and young adult writers. W-Pal is unique because it pioneered the combination of two powerful technology components—automated writing evaluation (AWE) and intelligent tutoring systems (ITSs). The theoretical and technological foundations of W-Pal (i.e., W-Pal's integration of AWE and intelligent tutoring) have been described in existing work (Crossley et al., 2016; Dai et al., 2010; Roscoe et al., 2014). AWE systems leverage natural language processing tools to evaluate student writing across several dimensions and provide feedback to students (Crossley et al., 2013; Grimes & Warschauer, 2010; Ranalli et al., 2017). One benefit of AWEs is that they can grade and offer individualized feedback to an entire classroom of students in a matter of seconds – a feat that would be impossible for a single teacher. Thus, AWEs offer scalability as well as opportunities for individual students to engage in more practice in less time. Existing evidence suggests that AWEs are effective when they provide opportunities for practice along with formative feedback that aligns with classroom instruction, which allows each student to receive support without additional strain on teachers' time and resources (Roscoe et al., 2011, 2013; Shute, 2008). Although AWE systems are helpful, their core purpose is to facilitate writing *assessment*. One limitation is that students may receive feedback that they do not know how to implement. Consequently, researchers and developers have investigated the extent to which AWE can be integrated into ITSs to provide students with *instruction* on the writing process (Roscoe et al., 2011; Roscoe & McNamara, 2013).

W-Pal incorporates elements of an ITS along with AWE to improve high school and college students' writing performance in the context of persuasive essays (e.g., Crossley et al., 2013; Roscoe & McNamara, 2013). Specifically, W-Pal provides feedback along with instructional modules that sequentially target writing *strategies*. Strategy instruction is provided to students via nine writing strategy modules that focus on particular strategies within one of the three phases of the writing process: prewriting, drafting, and revising. Prewriting modules include (a) freewriting and (b) planning. Drafting modules include (a) introduction building; (b) body building; and (c) conclusion building. The revising modules include (a) paraphrasing; (b) cohesion building; and (c) polishing of the text. An instructor or student can choose specific modules or complete each module in the default sequence. To provide strategy instruction, W-Pal uses animated pedagogical characters to explain writing strategies throughout a series of modules for prewriting, drafting, and revising. For each dimension of strategy instruction, the pedagogical character explains the strategy's purpose and meaning and then gives a concrete example of how to use the strategy, as well as a mnemonic to help students remember the strategy. For example, the conclusion-building module advises writers to *r*estate their thesis, *e*xplain how their thesis was supported, *c*lose the essay, *a*void new arguments, and *p*resent ideas in an interesting way (i.e., RECAP; Allen et al., 2016; Roscoe et al., 2015).

After each instructional module, W-Pal also presents opportunities to practice using the writing strategies. Specifically, each lesson is connected to game-based practice activities that help students to practice the strategy while also fostering engagement (Shank & Neeman, 2001; Taub et al., 2020) while students apply the writing strategies and reinforce the strategy knowledge acquired from W-Pal. As an example, students can play the *Essay Launcher* game to practice introduction building strategies. In this game, students try to rescue spaceships by selecting thesis statements and attention-grabbers for sample introduction paragraphs (see Figure 1). After students have practiced the strategies they learned from the instructional modules through game-based practice, they are given the opportunity to use the strategies in essay writing. Specifically, students



Lesson Game View (Press F11 to full screen or exit full screen)

INSTRUCTIONS

1. Use the Thesis Repair Control (bottom-right) to choose the thesis statement that best fits the paragraph.

2. Click to rotate the Technique Control wheel (bottom-left) to identify the technique used to write the paragraph.

3. Click Launch (top-left) when you are ready to launch your ship. The ship will malfunction if you make a mistake.

You will receive bonus points for extral fuel left over at the end of the mission!

Figure 1. Screenshot of a W-Pal Game.

practice writing timed persuasive essays that use SAT-style prompts. These prompts require no specialized knowledge but do require students to synthesize information and apply the strategies from the W-Pal modules.

When students submit training essays during their practice sessions, W-Pal's AWE system uses several text analysis tools (e.g., Coh-Metrix; Graesser & McNamara, 2012; Tool for the Automatic Analysis of Cohesion [TAACO], Crossley et al., 2016) to assess the student's essay for content, rhetorical style, text cohesion, and language sophistication (McNamara et al., 2013) and assign an overall quality rating from poor to great on a six-point scale. The essay is passed through a series of independent algorithms that drive feedback selection. First, the essay is checked for length. If the essay is too short, students receive feedback on idea development and content generation. Next, the system checks for structural elements of the essay (e.g., paragraphs). For example, if an essay consists of only two lengthy paragraphs, then the student receives feedback and instruction on how to structure an essay into an introduction, body, and conclusion. If these initial checks are passed, the introduction, body, and conclusion paragraphs are assessed individually via different algorithms. These paragraph-level algorithms use various linguistic indices to make inferences about paragraph quality. To illustrate, if the conclusion to an essay is flagged as low quality, the student receives strategy feedback and instruction on how to improve a conclusion. Thus, depending on the quality of individual sections in the submitted training essay, more targeted formative feedback may be provided for introduction, body, and/or conclusion building strategies. If an essay passes all basic thresholds, they receive feedback that encourages general revision and polishing. W-Pal's formative feedback, instructional modules, and writing practice have been shown to improve high school students' essay writing performance (Roscoe & McNamara, 2013). Additionally, students have reported that W-Pal's game-based practice is generally enjoyable (Allen et al., 2014; Roscoe & McNamara, 2013). Students have also shown greater gains in writing strategy knowledge after engaging with the game-based practice compared to traditional essay-based practice (Roscoe et al., 2014), with less skilled students showing greater gains than skilled students (Roscoe et al., 2019).

Although W-Pal provides writing practice and timely, appropriate feedback that is supplemented by instructional content, the system can be ineffective for certain learners. Some students simply ignore feedback messages and fail to adapt their writing (Wingate, 2010). Alternatively, students may *try* to thoughtfully consider the feedback, yet need additional practice and support to implement that strategy *effectively* in their writing. The original W-Pal was adaptive in that it provides tailored feedback on essays, but students proceeded through the system one instructional module at a time, and the selection of instructional content was made by the teacher or student. Thus, the original W-Pal system did not individualize instruction or tutoring—every student engaged with the same set of instructional modules in the same order. To increase the efficiency and effectiveness of the W-Pal system for improving students' writing performance, we sought to enhance the system by integrating adaptive just-in-time instruction that is individualized based on whichever instructional support each student needs most.

W-Pal Adaptive Instruction

The goal of redesigning W-Pal was to implement just-in-time instructional adaptivity in addition to the current individualized feedback that W-Pal provides. This adaptive instruction is akin to "outer-loop" adaptivity in that W-Pal uses the students' past performance (i.e., initial training essay) to direct them to different instructional modules (VanLehn, 2006). ITSs that include adaptive instruction demonstrate notable learning effects ($d = \sim 0.75$; VanLehn, 2011) over those that do not. Importantly, implementation of adaptive instruction in W-Pal means that students might no longer progress through W-Pal in the same linear sequence like they did in the traditional W-Pal. In fact, students may never even receive all of the instructional modules; instead, they receive only the instructional modules and practice opportunities that best align with their strengths and weaknesses.

More simply, one can imagine a one-on-one tutoring scenario. An instructor would prioritize some feedback and lessons over others. If a student's essay includes an effective introduction (e.g., with a thesis statement), then the instructor is unlikely to provide additional feedback or instruction on that topic. By contrast, if the instructor were reading the student's essay and noticed a weakness in its organization and cohesion, the instructor would not only provide specific feedback about that weakness but would also be more likely to provide additional *instruction* to support the student's understanding (given sufficient capacity to do so).

Procedurally, the new adaptive instruction is based on evaluations of students' essays to direct them to a specific strategy instruction module that targets a given weakness in the essay. The adaptive instruction thus further reinforces formative feedback. As shown in Figure 2, the student begins the "cycle" by writing the essay. W-Pal then evaluates the essay via a series of indices drawn from NLP tools (e.g., Coh-Metrix; McNamara & Graesser, 2012; TAACO, Crossley et al., 2016) that assess the essays for different categories of potential weaknesses associated with the writing strategies from the instructional modules (e.g., length, poor introduction), then the system provides feedback and directs the student to an instructional module based on this feedback. Table 1 provides a summary of feedback and instruction that targets five potential weaknesses in students' essays.

The adaptive instruction varies task selection based on student performance. For example, if a student's initial essay lacks a strong conclusion, then the student is directed to the conclusion module to watch the video lessons and

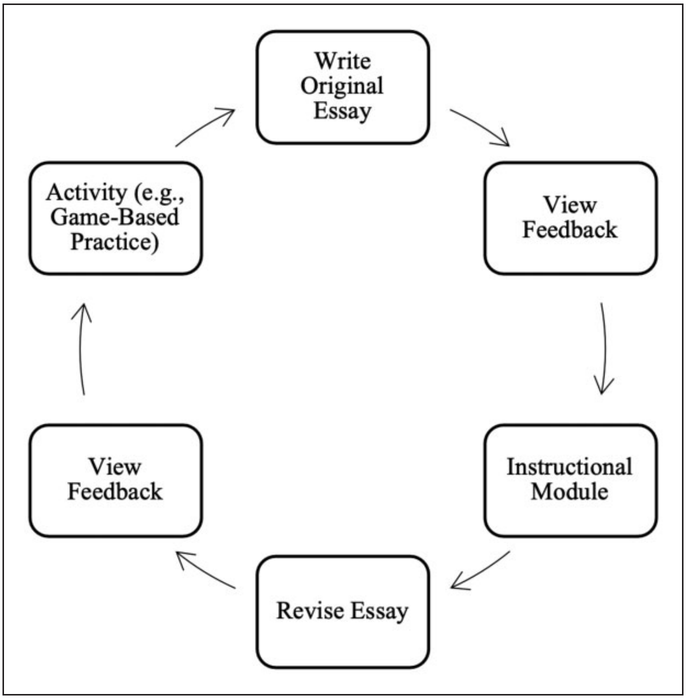


Figure 2. W-Pal Adaptive Cycle.

Table 1. Feedback and Instructional Targeting the Weakest Dimensions of Essay Quality.

Weakest dimension	Feedback topic	Instructional module
Length	Add more information	Freewriting
Structure	Clarify essay components	Planning
Introduction	Strengthen thesis statement	Introduction building
Body	Strengthen topic sentences	Body building
Conclusion	Summarize arguments	Conclusion building
General (no specific weakness)	Remove irrelevant information	Cohesion building; polishing

play games that emphasize strategies that correspond to building the conclusion. After the module has been completed, students then have the opportunity to revise their initial draft. The system assesses the revision and provides a second round of feedback. If the targeted weakness did not improve in the revised essay, then the system directs them to either watch the overview video for the corresponding instructional module or play games related to that module. If the revised essays have a different weakness than the original essay, the system directs the student to watch the overview video of the instructional module that corresponds to that weakness, which concludes the cycle for that session (see Figure 2). This adaptivity also results in a different instructional sequence for each student than a teacher would typically implement. That is, existing evidence shows that teachers mainly follow the sequence of prewriting, drafting, and revising, and as such, they tend to assign the modules linearly in the original W-Pal system (Roscoe & McNamara, 2013).

The Current Study

The purpose of the current study was to examine the extent to which W-Pal's newly implemented adaptive instruction would improve the quality of students' persuasive essays. Students completed pretest and posttest essays to gauge overall improvement over the course of their engagement with W-Pal. They engaged with the W-Pal adaptive system for four training sessions. During each training session, students wrote an initial training essay and were guided through different strategy instruction and games based on the automated evaluation of their initial essay in each session. Then, students had the opportunity to revise their training essay. Thus, we also examined incremental changes in the quality of students' essays within each of the four training sessions.

First, we examined the extent to which W-Pal improved essay quality from pretest to posttest, and more specifically, for whom improvement was greater. W-Pal's feedback and instruction was designed to improve writing performance particularly for students who struggle with literacy skills. In our analyses, we included reading comprehension skill as a moderator because existing research has shown that reading comprehension and writing skills are intimately connected via their shared demands on students' knowledge and cognitive skills (e.g., Allen et al., 2014). Thus, we aimed to examine whether W-Pal was more or less effective for students who had lower literacy skills. Overall, we hypothesized that the holistic quality of students' essays would improve from pretest to posttest. We also hypothesized that scores on the other essay quality dimensions that were targeted by the adaptive instruction (e.g., length, structure) would improve over the course of students' engagement with W-Pal.

Second, we examined the extent to which students' essays scores improved from the initial to revised versions for the four W-Pal training sessions. Because students received varied instruction for each of the four sessions to address the specific weaknesses in their initial training essays, we examined whether students improved on the dimensions for which they received instruction compared to the dimensions that were not directly targeted by instruction. We hypothesized that students' revised training essays would show greater improvement on the dimensions that received instruction relative to the dimensions that did not. Additionally, we expected that W-Pal's adaptive instruction would foster incremental improvements in the instructed dimensions of essay quality across the four training sessions. Thus, we also examined the extent to which students improved on the instructed dimensions across training sessions. We hypothesized that students' revised training essays would show overall improvement across the training sessions on the dimensions for which they received instruction, whereas they would not improve on the dimensions that were not targeted by instruction.

Method

Participants

Data were collected from $n = 56$ high school students from an urban school district in the southwestern United States (26 female, 29 male, 1 declined to answer). The mean age for students in the sample was 16.3 years ($SD = 1.3$ years, Range = 13–19 years). Self-reported demographic data indicated that the sample was 47% White, 31% Hispanic, 6% Black, 2% Asian, and 15% either selected *other* or did not report. Eighty-five percent of the sample reported English as their native language, 7% reported Spanish as their native language, and 8% reported other languages. Students were paid for their participation. Due to missing scores, data from five students were omitted from analyses.

Measures

System Log Data. Students' interactions with W-Pal were logged throughout the experiment. The resulting log data included: (a) each of the essays that students authored in W-Pal, (b) the feedback received on those essays, (c) the instructional modules that students were directed to after feedback on their initial essays, and (d) the revised essays after interacting with instructional modules. See Table 1 for examples of feedback categories, topics, and associated instructional modules.

Pretest and Posttest Essays. Students completed two timed (25 minute), SAT-style prompt-based essays at the beginning (Pretest Essay) and end of the experiment (Posttest Essay) in response to a prompt about either "competition" or "images." The "competition" prompt asked students to argue whether cooperation or competition leads to greater success; the "images" prompt asked students to argue whether images and impressions have too great an effect on decision making. The prompts were counterbalanced across students (i.e., half the students received the "images" prompt at pretest and the "competition" prompt at posttest, and the other half received the reverse order).

The pretest and posttest essays were evaluated with a scoring rubric that included a holistic score ranging from 1 (*very poor*) to 6 (*excellent*) and several quality dimensions (also 1–6) related to the introduction, body, conclusion, organization, cohesion, grammar, voice, word choice, and sentence structure (see Appendix for the holistic score rubric). Ratings were provided by two teams of two expert raters. The raters were doctoral students in English composition with over three years of experience teaching college writing and rating experience with standardized rubrics. The raters were first trained on the rubric using persuasive essays that were not included in this study. When raters reached an acceptable level of reliability (kappa >.70), they scored the persuasive essays such that two raters scored each essay. Scoring took place across two waves, each involving two raters. For each wave, raters independently scored the essays along each of the 10 components. For the first wave, reliability between the two raters ranged from acceptable to good across the four sub-scores (ICCs ranged from .71 to .82). For the second wave, the reliability between the two raters likewise ranged from acceptable to excellent between the two raters (ICCs ranged from .70 to .91). Ratings that differed by one point were averaged (i.e., a score of 2 by rater 1 and a score of 3 by rater 2 would result in a final score of 2.5). If any scores differed by more than one point, a third rater adjudicated the difference. However, pairs of human raters assigned scores that were in exact agreement or were within one point more than 95% of the time.

Reading Comprehension Skill. Students' reading comprehension skill was assessed using Form T of the Gates-MacGinitie Reading Test (GMRT; 4th ed.; MacGinitie & MacGinitie, 1989). The GMRT includes 48-item multiple-choice items to assess students' reading comprehension ability by asking students to read short passages and then answer two to six questions about the content of the passage designed to measure reading comprehension skill. The test was administered online via Qualtrics survey software. Existing research suggests that scores provide a reliable indicator of students' reading comprehension ability ($\alpha = .85$–$.92$; Phillips, Norris, Osmond, & Maynard, 2002). All students were given standard instructions, which included two practice questions. Students were given 20 minutes to answer as many questions as possible. The proportion of correct responses served as the outcome score.

Procedure

The study was conducted over a period of six days. All students completed pretest measures on the Day 1 of the study, including the GMRT and pretest essay. Then, students engaged with the W-Pal adaptive cycle for Days 2–5 (four training sessions). For each of these sessions, students wrote one persuasive training essay and then received feedback and an instructional module that targets the weakest dimension of their essay (i.e., length, introduction, body, conclusion, or structure; see Table 1 above). Then, all students were given 10 minutes to revise their essay. W-Pal then provided feedback on their revised essay. If W-Pal's AWE system indicated that the weakest dimension in the revised essay was the same as in the initial version, then students could choose to either watch the overview video for the corresponding instructional module or replay a practice game from that module. If the weakest dimension in the revision was different from the initial version, then students were directed to the overview video of the corresponding module. Students completed this cycle for each of the four training essays across the four training sessions. The four training essay prompts addressed different topics: loyalty, memories, patience, and winning. Finally, on Day 6, all students completed the posttest essay. The pretest and posttest essays did not include feedback, instruction, or revision opportunities.

Results

Preliminary Analysis

Each essay was assigned a holistic score that reflected the overall essay quality. To evaluate the validity of these scores, we conducted a principal components analysis on the pretest essays to establish whether holistic scores can serve as the primary measure of essay quality or whether we needed to separately analyze each of the dimension scores. The results revealed a single-component solution, $\chi^2(36) = 764.4$, $p < .001$, KMO = .920, that accounted for 68.6% of the total variance. We interpret this result as evidence that holistic scores indeed reflected overall essay quality as the synthesis of effective essay elements. The component score and the holistic essay scores at pretest were also very strongly correlated ($r = .93$, $p < .001$). Thus, subsequent analyses used holistic scores for overall

Table 2. Descriptives and Correlations Among Literacy Skill and Holistic Essay Rubric Scores.

Score	*M* (*SD*)	Range	1	2	3	4	5	6
Reading comprehension (GMRT prop. correct)	0.55 (.23)	.14–.94	1					
Pretest essay (Holistic)[a]	3.22 (.96)	1.0–5.5	.65***	1				
Training Essay 1 (Holistic)	3.30 (.81)	1.0–5.5	.58***	.52***	1			
Training Essay 2 (Holistic)	3.37 (.75)	2.0–5.0	.28*	.43***	.47***	1		
Training Essay 3 (Holistic)	3.42 (.82)	1.5–5.0	.46***	.49***	.50***	.49***	1	
Training Essay 4 (Holistic)	3.74 (.88)	1.5–6.0	.36**	.52***	.42***	.47***	.58***	1

Note. GMRT: Gates-MacGinitie Reading Test.

[a] Half of the participants' pretest essays were based on the "Images" prompt and half were based on the "Competition" prompt. Holistic scores did not differ between the "Images" prompt ($M = 3.2$, $SD = 1.0$) and "Competition" prompt ($M = 3.3$, $SD = 0.9$), $t(52) = .20$, $p = .84$.

*$p < .05$. **$p < .01$. ***$p < .001$.

essay quality. See Table 2 for descriptives and correlations among literacy skill (i.e., reading comprehension skill), the pretest essay scores, and initial training essay scores.

To What Extent Did W-Pal's Adaptive Instruction Improve Students' Essays from Pretest to Posttest?

We first examined the extent to which W-Pal's adaptive instruction fostered overall improvement from the *pretest* to *posttest*. In addition, we also investigated for whom W-Pal is beneficial. A goal of W-Pal is to improve writing performance, particularly for students with lower literacy skills. To examine improvement in essay quality from pretest to posttest, we conducted a series of separate repeated-measures ANCOVAs. In each model, instruction (pretest vs. posttest) was included as a within-subjects factor and literacy skills (i.e., reading comprehension skill) were included as a covariate. Each subscore (e.g., introduction, body, conclusion, and structure), as well as the holistic score were included as the dependent variables in their own respective models. Table 3 provides a summary of the ANCOVA results for the main effects of instruction (pretest vs. posttest) and literacy skills, as well as their interaction on students' scores for each essay dimension, as well as descriptive statistics of the essay dimension scores on the pretest essay and posttest essays.

As shown in Table 3, scores improved from pretest to posttest for the introduction, conclusion, and structure dimensions, whereas scores only marginally improved for the body. Importantly, there was a large overall improvement in holistic essay quality from pretest to posttest[1] (see Figure 3). Moreover, across all dimensions of essay quality, literacy skills were associated with performance,

Table 3. ANCOVA Results and Descriptive Statistics for Each Essay Dimension at Pretest and Posttest.

	Pretest	Posttest	Instruction	Literacy (GMRT)	Instruction × Literacy
	M (*SD*)	*M* (*SD*)	*F*-Ratio, *p*-value, effect size	*F*-Ratio, *p*-value, effect size	*F*-Ratio, *p*-value, effect size
Introduction	3.6 (0.9)	3.8 (0.8)	$F = 17.46$, p $< .001$, $\eta_p^2 = .26$	$F = 13.73$, $p < .001$, $\eta_p^2 = .22$	$F = 14.35$, $p < .001$, $\eta_p^2 = .23$
Body	3.5 (0.9)	3.7 (0.8)	$F = 3.50$, $p = .067$, $\eta_p^2 = .07$	$F = 23.07$, $p < .001$, $\eta_p^2 = .32$	$F = 1.28$, $p = .26$, $\eta_p^2 = .03$
Conclusion	3.0 (1.2)	3.2 (1.1)	$F = 4.75$, $p = .034$, $\eta_p^2 = .09$	$F = 29.37$, $p < .001$, $\eta_p^2 = .38$	$F = 3.86$, $p = .055$, $\eta_p^2 = .07$
Structure	3.6(0.8)	3.6 (0.7)	$F = 4.87$, $p = .032$, $\eta_p^2 = .09$	$F = 23.33$, $p < .001$, $\eta_p^2 = .32$	$F = 5.94$, $p = .019$, $\eta_p^2 = .11$
Holistic	3.2 (1.0)	3.5 (0.8)	$F = 8.16$, $p = .006$, $\eta_p^2 = .14$	$F = 35.50$, $p < .001$, $\eta_p^2 = .42$	$F = 4.99$, $p < .001$, $\eta_p^2 = .23$

GMRT: Gates-MacGinitie Reading Test.

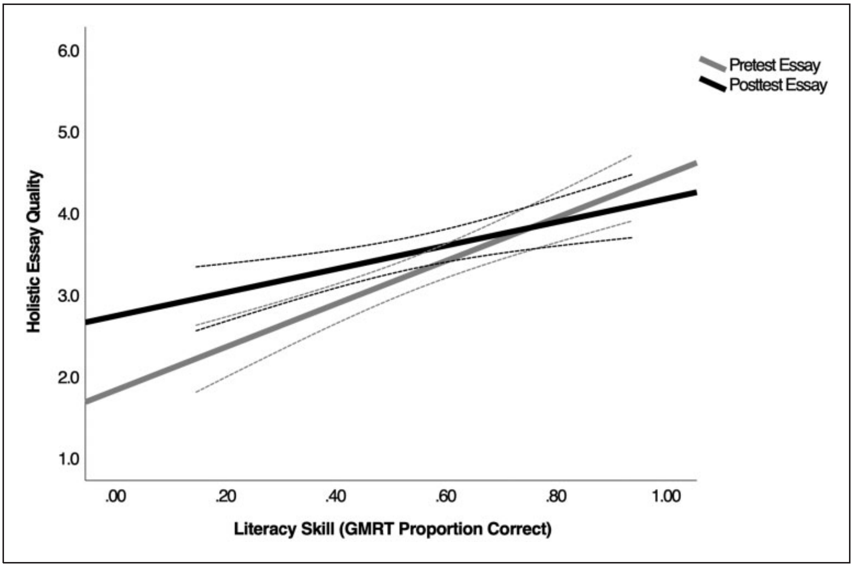


Figure 3. Improvement in Holistic Quality Prepost for High- and Low-Skilled Comprehenders. *Note.* Dashed lines represent 95% confidence intervals.

such that students with higher reading comprehension scores had higher scores overall. For the introduction, conclusion, structure, and holistic quality, students with lower literacy skills demonstrated greater improvement from pretest to posttest relative to students with higher literacy skills, whereas literacy skills did not influence improvement on the body of the essay.

These results suggest that students improved on both specific dimensions of essay quality and holistic quality from pretest to posttest. However, during engagement with the adaptive version of W-Pal, students received targeted strategy instruction on only the weakest of five dimensions (i.e., length, introduction, body, conclusion, structure) of their initial training essays within each of the training sessions. Therefore, it is also critical to examine the extent to which students showed more specific improvements on the dimensions for which they received targeted instruction.

To What Extent Did W-Pal's Adaptive Instruction Improve Weaknesses in Students' Training Essays from the Initial to Revised Versions?

Because students received feedback and instruction on the weakest dimension of their initial training essays, we first used log data to identify which dimension of each student's initial training essay tended to show weaknesses for each session. Students varied in the dimensions they struggled with across the four essays.

Table 4. Percent of Students Receiving Instruction Modules Targeting the Weakest Essay Dimensions.

		Training session				
Weakest dimension	Instructional module	One	Two	Three	Four	Avg.
Length/elaboration	Freewriting	43%	57%	36%	36%	43%
Structure	Planning	9%	4%	4%	2%	10%
Introduction	Introduction building	20%	16%	36%	38%	28%
Body	Body building	0%	2%	5%	2%	2%
Conclusion	Conclusion building	13%	7%	13%	7%	10%
General (no specific weakness)	Paraphrasing/polishing	16%	13%	7%	13%	12%

Table 4 shows the percentage of students in the sample who received strategy instruction targeting the dimensions of essay quality across the four training essays. The dimension that most frequently received targeted instruction was length. Students who were weakest on length were directed to the freewriting instructional module, as insufficient essay length indicated a lack of elaboration and idea development. The freewriting module encouraged idea development and could therefore have fostered improvement in any or all of the other essay dimensions (i.e., students could increase length of the introduction, body, conclusion, or some combination of the three), but it most strongly emphasized substantiating the body of the essay. The second most frequently instructed dimension was the introduction. Students who were weakest in structure, introduction, body, and conclusion were directed to their respective instructional modules. Note that students could have also received "general" writing instruction aimed at increasing completeness and clarity only when their essay passed all of W-Pal's thresholds for the other dimensions.

To examine the extent to which the adaptive instruction improved scores on students' training essays from the initial to revised versions, we used system log data to first isolate the scores for the weakest dimension in the initial and revised version of each training essay. Then, we compared each student's isolated scores to the mean score for the dimensions that were not directly targeted by instruction. For example, if W-Pal determined that a student's weakest feature in their initial training essay was the conclusion, then W-Pal would direct that student to the conclusion building module prior to revising. For this student, we would examine improvement on the *conclusion* relative to the other essay dimensions. Note that our operationalization of "un-instructed" dimensions does not imply that the strategy instruction could not also have improved additional dimensions, as they are inherently interconnected (Allen et al., 2014). Instead, we mean that these dimensions were not the direct target of the instruction.

Because each student could have received different instructional modules for each training essay, we compared improvement in the instructed dimensions relative to the uninstructed dimensions for each essay. Thus, we conducted a series of four separate repeated-measures ANOVAs that included essay version (initial vs. revised) and instruction (instructed vs. un-instructed) as within-subjects factors and scores on the instructed and un-instructed respective training essays as the dependent variables in their respective models.

For Essay 1, there was a main effect of version, $F(1, 50) = 5.04$, $p = .029$, $\eta_p^2 = .09$, such that scores improved from the initial ($M = 3.46$, $SE = .11$) to revised version ($M = 3.74$, $SE = .09$). This was qualified by a Version × Instruction interaction, $F(1, 50) = 9.08$, $p = .004$, $\eta_p^2 = .15$ (see Figure 4), such that the instructed dimensions improved from the initial ($M = 3.35$, $SE = .13$) to revised version ($M = 3.75$, $SE = .11$), whereas the un-instructed dimension did not differ between the initial ($M = 3.56$, $SE = .10$) to revised version ($M = 3.72$, $SE = .09$). No other effects reached significance. For Essay 2, there was a main effect of instruction, $F(1, 49) = 10.98$, $p < .001$, $\eta_p^2 = .21$, such that scores were higher overall on the uninstructed dimensions ($M = 3.74$, $SE = .06$) than on the instructed dimensions ($M = 3.56$, $SE = .09$). The Version × Dimension interaction did not reach significance. It is critical to emphasize that the instructed dimensions were also the weakest dimensions, so it is unsurprising that the scores for these dimensions were lower overall. Likewise, for Essay 3, there was also a moderate main effect of instruction,

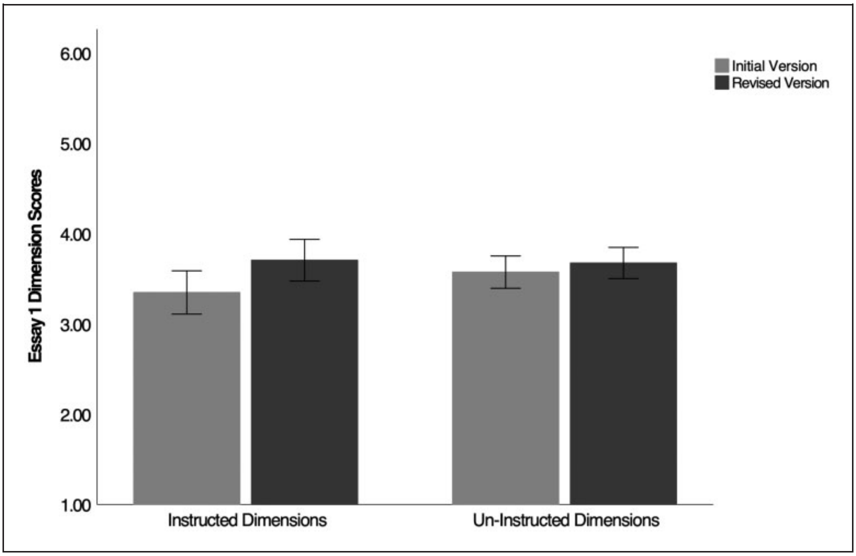


Figure 4. Improvement on Instructed Versus Un-Instructed Dimensions for Training Essay 1.
Note. Error bars represent 95% confidence intervals.

$F(1, 55) = 5.29$, $p = .025$, $\eta_p^2 = .09$, such that scores were higher on the un-instructed dimensions ($M = 3.68$, $SE = .06$) than on the instructed dimension ($M = 3.55$, $SE = .08$). The Version × Dimension interaction did not reach significance. For Essay 4, no main effects or interactions reached significance.

These results show that students' training essays improved on the instructed dimensions relative to the un-instructed dimensions in some cases (i.e., Essay 1). It is also possible that W-Pal's adaptive instruction led to incremental improvements in the instructed dimensions of essay quality *across* the four training sessions.

To examine this, we conducted two repeated-measures ANOVAs with session (sessions 1 - 4) and version (initial vs. revised) as within-subjects variables. For the first model, the dependent variable was the rubric scores on instructed dimensions; for the second model, the dependent variable was the rubric scores on un-instructed dimensions.

For the first model, there was a main effect of session, $F(3, 156) = 3.34$, $p = .021$, $\eta_p^2 = .06$, such that essay scores were higher for session 4 ($M = 3.78$, $SE = .09$) than for session 1 ($M = 3.54$, $SE = .10$), session 2 ($M = 3.54$, $SE = .09$), and session 3 ($M = 3.56$, $SE = .09$), $ps < .01$. There was also a Session × Version interaction, $F(3, 156) = 2.92$, $p = .036$, $\eta_p^2 = .05$, such that scores significantly improved for session 1 from the initial ($M = 3.36$, $SE = .13$) to revised essay version ($M = 3.79$, $SE = .11$), $p < .001$. Scores did not differ between the initial and revised versions for the remaining essays ($ps > .10$).

For the second model, there was a main effect of session, $F(3, 156) = 2.85$, $p = .040$, $\eta_p^2 = .05$, such that essay scores were higher for session 4 ($M = 3.86$, $SE = .06$) than for session 1 ($M = 3.67$, $SE = .07$) and session 3 ($M = 3.72$, $SE = .07$), $ps < .05$, but did not differ from session 2 ($M = 3.74$, $SE = .06$). This was qualified by a Session × Version interaction, $F(3, 156) = 3.00$, $p = .032$, $\eta_p^2 = .06$, such that revised scores on Essay 4 ($M = 3.72$, $SE = .09$) were *lower* than initial scores ($M = 3.94$, $SE = .09$), whereas scores for Essays 1 to 3 did not differ between the initial and revised versions ($ps > .10$).

These results show that students' scores on the instructed dimensions significantly improved from the initial to revised versions on Essay 1, but not on the others. By contrast, students did not significantly improve on the un-instructed dimensions in any of the sessions. Students also showed improvement overall from the first session to the fourth session. Overall, these results suggest that adaptive instruction may benefit the quality of students' essays in the training sessions, as well as from pretest to posttest.

Discussion

The goal of the current study was to examine the extent to which W-Pal's adaptive strategy instruction would lead to improvements in high school students' writing. Adaptive instruction aims to provide students with a more *individualized* intelligent tutoring experience by administering writing instruction, practice, and feedback that targets each student's specific weaknesses. The new adaptive instruction version of W-Pal was designed to address students' most critical needs and was therefore expected to support improvements in essay quality, particularly on dimensions with which students had the most difficulty.

We first examined overall improvement in essay quality from pretest to posttest and also whether these improvements were moderated by reading comprehension skill, as W-Pal's feedback and instruction was intended to help students with lower literacy skills. The results showed that holistic essay quality improved from pretest to posttest only for students with lower literacy skills. Thus, W-Pal seems to be beneficial when used by its target audience. Next, we examined the dimensions of essay quality for which students could have received instruction across their engagement with W-Pal. The results revealed that scores for the introduction, conclusion, and structure improved markedly from pretest to posttest, with greater improvements amongst students with lower reading comprehension skills. These findings generally supported our hypothesis that W-Pal can lead to improvements in essay quality. In turn, this suggests that W-Pal fostered writing strategy uptake overall, especially for students who need writing strategies the most (Allen et al., 2014). It is promising that W-Pal improves writing quality overall, but it is also critical to consider improvements to essay quality at a finer grain size.

Specifically, the redesigned W-Pal varied instruction for each student depending on the weakness in each of the initial training essays. For each essay, students could receive formative feedback and instruction to improve weaknesses in essay length, introduction, body, conclusion, or structure, as well as general instruction to increase clarity and completeness. Log data indicated that the most frequently administered instruction was freewriting (43% of students) to encourage idea development and elaboration. Instruction on freewriting could have improved any dimension of essay quality depending on where students chose to add more content, but the freewriting instruction most emphasized the essay body. The next most frequent instruction was aimed at improving the introduction (28% of students). Thus, there was variability with respect to the strategy instruction students received across the training sessions.

We next examined the effectiveness of the adaptive instruction for improving dimensions of essay quality for each training essay. We hypothesized that students' scores would improve from the initial to revised version on the dimensions that were targeted by instruction relative to those that were not. The results indicated that scores improved on the dimensions that received instruction relative to those that did not for Essay 1. For training Essays 2 and 3, scores on the instructed dimensions did not improve and were lower than scores on the un-instructed dimensions. This is unsurprising, because W-Pal provided instruction only for the weakest (i.e., lowest scoring) dimension. Overall, these results provided partial support for our hypothesis.

We also examined the extent to which students' essay scores on the instructed dimensions of essay quality improved across the four training sessions. It is possible that the adaptive instruction could have yielded incremental improvement in the instructed dimensions, but we would expect relatively less improvement on the dimensions that were not directly targeted by instruction. Accordingly, we hypothesized that essay scores would improve overall on the instructed dimensions, but not for the un-instructed dimensions. However, our results did not support this hypothesis. The results suggested that scores on the instructed dimensions did not improve overall across the four sessions, but as before, session 1 scores improved from the initial to revised essay version. Scores on uninstructed dimensions did not improve from the initial to revised version, nor did they improve on any of the training essays.

There are several possible explanations for the limited improvement on instructed dimensions of essay quality across the four training sessions. Specifically, it may be the case that the strategy instruction went beyond the targeted dimension and influenced other dimensions of essay quality. In the present study, this would mean that the instruction also influenced scores on the "un-instructed" dimensions. This possibility is plausible given evidence that aspects of writing are interconnected (Allen et al., 2014). Although such instructional spillover may be helpful for students' overall writing performance, it would create difficulty in detecting the effect of instruction on any particular dimension of essay quality. It is also possible that there were essay-specific demands that influenced how amenable particular dimensions were to improvement during revision (Allen et al., 2016). Indeed, the results indicated that, for some training essays, there was no improvement from the initial to revised versions.

These results also highlight interesting connections between students' improvements in the training essays and their overall improvements from pretest to posttest. First, the largest improvement from pretest to posttest was on the introduction. Log data from the training sessions indicated that the introduction was a frequent weakness in students' initial training essays and was thus frequently targeted by strategy instruction. It may be the case that the overall improvement observed for the introduction from pretest to posttest was due to the frequency of instruction that targeted the introduction. Second, students also improved markedly in holistic essay quality from pretest to posttest. The log data also indicated that the most common instruction was for idea development and elaboration (i.e., the freewriting module). Although the freewriting module emphasized substantiating the body of the essay, the instruction was relatively nonspecific compared to the other modules. Thus, the freewriting instruction could have improved essay quality on any dimension, as students were free to elaborate on content anywhere in the essay during revision. These additions could have substantially contributed to improvements in holistic quality. However, these are merely speculations, as we do not have data that directly supports these conclusions.

The findings of the present study should be interpreted in light of its limitations. First, we had a small sample size, which imposed limitations in statistical power. Second, the current study did not include a control group comparison that used a pre-sequenced version. Ideally, the new version of W-Pal with adaptive instruction would outperform the standard W-Pal in terms of improvements to writing performance, but we were unable to provide such evidence. Third, the instructional period lasted for only four training sessions over four days. Thus, students only produced four training essays. We would expect the effects observed in the current study to be amplified over a longer instructional period with more opportunities for practice and targeted feedback and instruction. Thus, future studies should compare W-Pal's adaptive instruction to the standard W-Pal with a larger sample over a longer instructional period. Doing so would provide a much more robust test of the benefits of adaptive writing strategy instruction.

Despite the inherent limitations of this study, it is critical to emphasize the importance of these results given the frequent lack of uptake of feedback on students' essays (e.g., Carless & Boud, 2018), and the vital importance of providing students with just-in-time instruction that focuses on their individual needs. Students cannot improve their writing if they are merely made aware of their weaknesses; rather, they must be provided with specific, formative feedback that provides them with suggestions on how to *improve* their essays. Further, students benefit from instruction that exemplifies the strategies and provides opportunities to practice applying them (i.e., game-based practice). This study provides preliminary evidence that the combination of targeted writing strategy instruction along with essay writing practice can improve the quality of students' writing, particularly for students who tend to struggle with literacy skills. The efficacy of automated adaptive instruction is critical because teachers' limited time and resources makes it impossible for them to develop and deliver effective personalized instruction and feedback on students' writing. Thus, W-Pal (and other writing-focused ITSs) may complement classroom instruction by supplying additional practice and more individualized feedback and instruction. Overall, the results of this study indicate that the combination of feedback and instruction on writing strategies can provide an efficient and effective means of improving students' writing, without placing additional burdens on educators.

Appendix: Holistic Rating Form

After reading each essay and completing the analytical rating form, assign a holistic score based on the rubric below. For the following evaluations you will need to use a grading scale between 1 (minimum) and 6 (maximum). As with the analytical rating form, the distance between each grade (e.g., 1–2, 3–4, 4–5) should be considered equal.

SCORE OF 6: An essay in this category demonstrates clear and consistent mastery, although it may have a *few minor errors*. A typical essay effectively and insightfully develops a point of view on the issue and demonstrates outstanding critical thinking, using clearly appropriate examples, reasons, and other evidence to support its position is well organized and clearly focused, demonstrating clear coherence and smooth progression of ideas exhibits skillful use of language, using a varied, accurate, and apt vocabulary demonstrates meaningful variety in sentence structure is free of most errors in grammar, usage, and mechanics.

SCORE OF 5: An essay in this category demonstrates reasonably consistent mastery, although it will have *occasional errors or lapses in quality*. A typical essay effectively develops a point of view on the issue and demonstrates strong critical thinking, generally using appropriate examples, reasons, and other evidence to support its position is well organized and focused, demonstrating coherence and progression of ideas exhibits facility in the use of language, using appropriate vocabulary demonstrates variety in sentence structure is generally free of most errors in grammar, usage, and mechanics.

SCORE OF 4: An essay in this category demonstrates adequate mastery, although it will have *lapses* in quality. A typical essay develops a point of view on the issue and demonstrates competent critical thinking, using adequate examples, reasons, and other evidence to support its position is generally organized and focused, demonstrating some coherence and progression of ideas exhibits adequate but inconsistent facility in the use of language, using generally appropriate vocabulary demonstrates some variety in sentence structure has some errors in grammar, usage, and mechanics.

SCORE OF 3: An essay in this category demonstrates developing mastery, and is *marked* by ONE OR MORE of the following weaknesses: develops a point of view on the issue, demonstrating some critical thinking, but may do so inconsistently or use inadequate examples, reasons, or other evidence to support its position is limited in its organization or focus, or may demonstrate some lapses in coherence or progression of ideas displays developing facility in the use of language, but sometimes uses weak vocabulary or inappropriate word choice lacks variety or demonstrates problems in sentence structure contains an accumulation of errors in grammar, usage, and mechanics.

SCORE OF 2: An essay in this category demonstrates little mastery, and is *flawed* by ONE OR MORE of the following weaknesses: develops a point of view on the issue that is vague or seriously limited, and demonstrates weak critical thinking, providing inappropriate or insufficient examples, reasons, or other evidence to support its position is poorly organized and/or focused, or demonstrates serious problems with coherence or progression of ideas displays very little facility in the use of language, using very limited vocabulary or incorrect word choice demonstrates frequent problems in sentence structure contains errors in grammar, usage, and mechanics so serious that meaning is somewhat obscured.

SCORE OF 1: An essay in this category demonstrates very little or no mastery, and is *severely flawed* by ONE OR MORE of the following weaknesses: develops no viable point of view on the issue, or provides little or no evidence to support its position is disorganized or unfocused, resulting in a disjointed or incoherent essay displays fundamental errors in vocabulary demonstrates severe flaws in sentence structure contains pervasive errors in grammar, usage, or mechanics that persistently interfere with meaning.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by IES Grants R305A180261 and R305A180144, as well as the Office of Naval Research (Grant: N00014-17-1-2300). Opinions, conclusions, or recommendations do not necessarily reflect the view of the Department of Education, IES, or the Office of Naval Research.

ORCID iD

Reese Butterfuss https://orcid.org/0000-0001-9326-4176

Note

1. The most frequently instructed dimension of essay quality was length, which indicates a lack of idea development. Thus, we examined the extent to which increases in length accounted for improvements in holistic essay quality from pretest to posttest. Indeed, students' essays increased in length (i.e., contained more words) from the pretest essay ($M = 307$, $SD = 129$) to the posttest essay ($M = 338$, $SD = 107$), $t(52) = 2.22$, $p = .030$. However, including change in essay length from pretest to posttest as a covariate in the analysis of holistic essay scores revealed that the increase in length did not account for improvements in holistic scores, $F(1, 47) = 0.13$, $p = .72$. Thus, improvements in essay quality from pretest to posttest were not merely due to longer essays at posttest.

References

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). Computer-based writing instruction. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 316–329). Guilford Press.

Allen, L. K., Likens, A. D., & McNamara, D. S. (2019). Writing flexibility in argumentative essays: A multidimensional analysis. *Reading and Writing*, *32*(6), 1607–1634.

Allen, L. K., Snow, E. L., Jackson, G. T., Crossley, S. A., & McNamara, D. S. (2014). Reading components and their relation to writing. *Topics in Cognitive Psychology*, *114*(4), 663–691.

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325.

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, *11*(vol. 11 issue 3), 415–443.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, *32*, 1–16.

Crossley, S. A., Roscoe, R., & McNamara, D. S. (2014). What is successful writing? An investigation into the multiple ways writers can write successful essays. *Written Communication*, *31*(2), 184–214.

Crossley, S., Varner, L. K., Roscoe, R. D., & McNamara, D. S. (2013). Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial intelligence in education* (pp. 269–278). Springer. http://link.springer. com/chapter/10.1007/978-3-642-39112-5_28

Dai, J., Raine, R. B., Roscoe, R., Cai, Z., & McNamara, D. S. (2010). The Writing-Pal tutoring system: Development and design. *Journal of Engineering and Computer Innovations*, *2*(1), 1–11.

Dempsey, M. S., PytlikZillig, L. M., & Bruning, R. H. (2009). Helping preservice teachers learn to assess writing: Practice and feedback in a web-based environment. *Assessing Writing*, *14*(1), 38–61.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406.

Flower, L. S., & Hayes, J. R. (1980). The cognition of discovery: Defining a rhetorical problem. *College Composition and Communication*, *23*, 2–32.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365–387.

Graesser, A. C., & McNamara, D. S. (2012). Automated analysis of essays and open-ended verbal responses. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics* (pp. 307–325). American Psychological Association.

Graham, S., Capizzi, A., Harris, K. R., Hebert, M., & Morphy, P. (2014). Teaching writing to middle school students: A national survey. *Reading and Writing*, *27*(6), 1015–1042.

Graham, S., & Hall, T. E. (2016). Writing and writing difficulties from primary grades to college. *Learning Disability Quarterly*, *39*(1), 3–4.

Graham, S., Kiuhara, S. A., & MacKay, M. (2020). The effects of writing on learning in science, social studies, and mathematics: A meta-analysis. *Review of Educational Research*, *90*(2), 179–226.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, *99*(3), 445–476.

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment*, *8*, 4–43.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1–27). Erlbaum.

Johnstone, K. M., Ashbaugh, H., & Warfield, T. D. (2002). Effects of repeated practice and contextual-writing experience on college students' writing skills. *Journal of Educational Psychology*, *94*(2), 305–315.

Kellogg, R. T., & Raulerson, B. A. (2007). Improving the writing skills of college students. *Psychonomic Bulletin & Review*, *14*(2), 237–242.

Kellogg, R. T., & Whiteford, A. P. (2009). Training advanced writing skills: The case for deliberate practice. *Educational Psychologist*, *44*(4), 250–266.

Kieft, M., Rijlaarsdam, G., Galbraith, D., & Bergh, H. (2007). The effects of adapting a writing course to students' writing strategies. *British Journal of Educational Psychology*, *77*(3), 565–578.

Kiuhara, S. A., Graham, S., & Hawken, L. S. (2009). Teaching writing to high school students: A national survey. *Journal of Educational Psychology*, *101*(1), 136–160.

MacGinitie, W. H., & MacGinitie, R. K. (1989). *Gates–MacGinitie reading tests*. Riverside.

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). IGI Global. https://doi.org/10.4018/978-1-60960-741-8

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*(2), 499–515.

Miller, D., Scott, C., & McTigue, E. (2018). Writing in secondary-level disciplines: A systematic review of context, cognition, and content. *Educational Psychology Review*, *30*(1), 83–120.

National Commission on Writing (2004). *Writing: A ticket to work. Or a ticket out*. College Board.

Phillips, L. M., Norris, S. P., Osmond, W. C., & Maynard, A. M. (2002). Relative reading achievement: A longitudinal study of 187 children from first through sixth grades. *Journal of Educational Psychology*, *94*(1), 3–13.

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, *37*(1), 8–25.

Roscoe, R. D., Allen, L. K., & McNamara, D. S. (2019). Contrasting writing practice formats in a writing strategy tutoring system. *Journal of Educational Computing Research*, *57*(3), 723–754.

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.

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*(4), 1010–1025.

Roscoe, R. D., Snow, E. L., Allen, L. K., & McNamara, D. S. (2015). Automated detection of essay revising patterns: Application for intelligent feedback in a writing tutor. *Technology, Instruction, Cognition, and Learning*, *10*, 59–79.

Roscoe, R. D., Varner, L. K., Cai, Z., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2011). Internal usability testing of automated essay feedback in an intelligent writing tutor. In R. C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th international Florida artificial intelligence research society (FLAIRS) conference* (pp. 543–548). AAAI Press.

Roscoe, R. D., Varner, L. K., Crossley, S. A., & McNamara, D. S. (2013). Developing pedagogically-guided algorithms for intelligent writing feedback. *International Journal of Learning Technology*, *8*(4), 362–381.

Shank, R., & Neeman, A. (2001). Motivation and failure in educational systems design. In K. Forbus & P. Feltovich (Eds.), *Smart machines in education* (pp. 37–69). AAAI Press/MIT Press.

Shute, V. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189.

Taub, M., Sawyer, R., Smith, A., Rowe, J., Azevedo, R., & Lester, J. (2020). The agency effect: The impact of student agency on learning, emotions, and problem-solving behaviors in a game-based learning environment. *Computers & Education*, *147*, 103781–103719.

Torrance, M., Thomas, G. V., & Robinson, E. J. (2000). Individual differences in undergraduate essay-writing strategies: A longitudinal study. *Higher Education*, *39*(2), 181–200.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence and Education*, *16*, 227–265.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221.

Wingate, U. (2010). The impact of formative feedback on the development of academic writing. *Assessment & Evaluation in Higher Education*, *35*(5), 519–533.

Author Biographies

Reese Butterfuss is a postdoctoral research scholar at Arizona State University. His research focuses on the cognitive processes that underlie learning from texts, as well as improving students' literacy skills using technology-based literacy instruction.

Rod D. Roscoe is an associate professor of Human Systems Engineering, and a Diane and Gary Tooker Endowed Professor for Effective Education in STEM at Arizona State University. His research investigates the intersection of learning science, computer science, and user science in educational technology design, and how this integration can inform effective and innovative uses of such technologies. He also studies the role and impact of human-centered instruction in engineering education.

Laura K. Allen is an assistant professor of Psychology at University of New Hampshire. Her research examines the cognitive and metacognitive processes involved in learning from text, and she applies those insights to educational practice through the development of interventions and educational technologies.

Kathryn S. McCarthy is an assistant professor of educational psychology at Georgia State University. Her research explores the higher-order processes involved in reading comprehension and how these processes vary as a function of discipline and readers' individual differences. She also examines how both in-person interventions and educational technology can be leveraged to support successful learning from text.

Danielle S. McNamara is a professor in the Psychology Department at Arizona State University. She focuses on educational technologies and discovering new methods to improve students' ability to understand challenging text, learn new information, and convey their thoughts and ideas in writing. Her work integrates various approaches and methodologies including the development of game-based, intelligent tutoring systems (e.g., iSTART, Writing Pal), the development of natural language processing tools (e.g., iSTART, Writing Pal, CohMetrix, the Writing Assessment Tool), basic research to better understand cognitive and motivational processes involved in comprehension and writing, and the use of learning analytics across multiple contexts. More information about her research and access to her publications are available at soletlab.com.