Journal of Educational Psychology

Writing Pal: Feasibility of an Intelligent Writing Strategy Tutor in the High School Classroom

Rod D. Roscoe and Danielle S. McNamara

Online First Publication, September 9, 2013. doi: 10.1037/a0032340

Journal of Educational Psychology

0022-0663/13/$12.00 DOI: 10.1037/a0032340

Writing Pal: Feasibility of an Intelligent Writing Strategy Tutor in the High School Classroom

Rod D. Roscoe and Danielle S. McNamara
Arizona State University

The Writing Pal (W-Pal) is a novel intelligent tutoring system (ITS) that offers writing strategy instruction, game-based practice, essay writing practice, and formative feedback to developing writers. Compared to more tractable and constrained learning domains for ITS, writing is an ill-defined domain because the features of effective writing are difficult to quantify and individual writers can employ diverse strategies to achieve similar goals. The development of an ITS in an ill-defined domain presents particular challenges regarding comprehensive instruction, modularized content, extended practice, and formative feedback. In this article, we describe how the development of W-Pal has uniquely addressed these concerns and present the results of a study assessing the feasibility of this system in high school English classrooms. This study included 2 teachers and their 141 10th grade English class students who utilized W-Pal over a 6-month period during the academic year. Log-file analyses showed that students used all aspects of W-Pal, but activity and engagement was uneven throughout the year and decreased over time. Essay scores improved over time and surveys indicated that students perceived the lessons, games, and feedback as beneficial. However, specific aspects of the learning environment were critiqued as annoying, challenging, or lacking specificity. Overall, the results suggest that the system was generally well-received by the students but also offer insights for the development of ITSs in ill-defined domains.

Keywords: intelligent tutoring systems, writing instruction, usability and feasibility testing, ill-defined learning domains

Intelligent tutoring systems (ITSs) provide adaptive, interactive, computer-based support for learning based on sound pedagogical principles (Graesser, McNamara, & VanLehn, 2005), and educators now have access to effective intelligent tutors in domains such as mathematics (Beal, Arroyo, Cohen, & Woolf, 2010), geometry (Aleven & Koedinger, 2002), biology (Michael, Rovick, Glass, Zhou, & Evens, 2003), physics (Graesser et al., 2004; VanLehn et al., 2005), computer literacy (Graesser et al., 2004), reading comprehension (McNamara, O'Reilly, Best, & Ozuru, 2006), and foreign language (Gamper & Knapp, 2002; Johnson & Wu, 2008). In this study, we examine the Writing Pal (W-Pal), an ITS that offers *writing strategy instruction* along with game-based practice, essay writing practice, and formative feedback to high school students. Historically, ITS development has focused on well-defined learning domains, in which fundamental concepts, procedures, and evaluation criteria are relatively constrained. In contrast, writing is an *ill-defined learning domain* because the features of skilled writing are difficult to quantify, and individual writers may employ diverse strategies to achieve similar goals.

A particular focus of this study is how high school students perceive intelligent tutoring of writing in the classroom (Grimes & Warschauer, 2010). For ill-defined domains, in which evaluations of students' work are inherently debatable, such subjective reactions are crucial. Students who rebuff the ITS are unlikely to engage with the system over meaningful periods of instruction (i.e., several weeks, a semester, or a school year). Thus, we assume that feasibility depends upon whether the system is perceived as valid and valuable. At this stage in W-Pal's development, an experimental test of instructional efficacy was not warranted. Rather, it was most important for us to examine a) how and whether students use the W-Pal over time and b) students' perceptions of the utility and design of W-Pal. These data help to define the feasibility of the system and inform later development and deployment.

Computer Support for Writing Instruction

Several technologies have been developed to support students' writing by grading essays (Grimes & Warschauer, 2010; Shermis & Burstein, 2003), teaching summarization (Kintsch, Caccamise, Franzke, Johnson, & Dooley, 2007) and argumentation skills (Wolfe, Britt, Petrovich, Albrecht, & Kopp, 2009), or scaffolding essay composition (Proske, Narciss, & McNamara, 2012; Rowley & Meyer, 2003). An important question is how well technologies address the pedagogical needs arising from the ill-defined nature of writing. Ill-structured problems possess ambiguous goals, solution paths, or assessment criteria (Simon, 1973). Lynch, Ashley, Pinkwart, and Aleven (2009, p. 258) argued that learning domains

Rod D. Roscoe and Danielle S. McNamara, Learning Sciences Institute, Arizona State University.

The research reported here was supported by Institute of Education Sciences, U.S. Department of Education Grant R305A080589 to Arizona State University. The opinions expressed are those of the authors and do not necessarily represent views of the institute or the U.S. Department of Education.

Correspondence concerning this article should be addressed to Rod D. Roscoe, Learning Sciences Institute, Arizona State University, PO Box 872111, Tempe, AZ 85287-2111. E-mail: rod.roscoe@asu.edu

are ill-defined when "essential concepts, relationships, and procedures for the domain" and the "means to validate problem solutions or cases" are not specified by a single strong domain theory. There may be multiple conceptualizations of key problems and tasks and there may be multiple approaches for solving those problems. Given such ambiguity, assessment of solutions may also be context-dependent and subjective. Thus, ITSs in ill-defined domains must not only address the challenges common to any educational technology, but must also overcome unique hurdles that arise when appropriate content, tasks, evaluation, and feedback are uncertain.

The ill-defined nature of writing emerges from the many nonlinear and interactive tasks that comprise the writing process (Deane et al., 2008; Flower & Hayes, 1981). For example, *prewriting* involves generating and organizing ideas prior to writing, and *drafting* involves translating initial ideas and plans into coherent text. In persuasive writing, writers must frame their arguments precisely and objectively and support arguments with factual evidence. Subsequently, *revising* entails elaborating and reorganizing the text to improve overall quality. Throughout these stages, writers also develop cohesion, style, voice, and other global qualities. To help students navigate these complex demands, writing pedagogy emphasizes the importance of *strategy instruction* that equips students with (a) concrete strategies for diverse writing processes, (b) background knowledge for using the strategies, and (c) opportunities for extended practice (Graham, McKeown, Kiuhara, & Harris, 2012; Graham & Perin, 2007). Effective interventions teach explicit strategies for planning, drafting, editing, and summarizing, along with information about how and why the strategies should be used (De la Paz & Graham, 2002).

Another aspect of the ill-defined nature of writing is the subjectivity of evaluation. Every essay exhibits unique content and errors that represent individual students' writing processes. To assign a score, essay graders (e.g., teachers) must interpret the appropriateness of these decisions in the context of the assignment. Writing assessment research has found this process to be challenging (Huot, 1996; Meadows & Billington, 2005). Over time and multiple instances of grading, human graders are unlikely to assign the same grades to the same essays consistently unless carefully trained to do so (Crossley & McNamara, 2011; Meadows & Billington, 2005). Such subjectivity also raises questions about how to give meaningful feedback. Research has emphasized the importance individualized, formative feedback that describes clear methods for improvement (McGarrell & Verbeem, 2007; Shute, 2008), such as strategies for developing arguments and evidence. In contrast to summative feedback on overall performance, formative feedback supports writing proficiency by making the means of progress explicit.

An analysis of writing instruction from the perspective of ill-defined learning domains thus suggests several design principles that are germane to any writing ITS. An intelligent writing tutor may need to combine (a) comprehensive strategy instruction across multiple phases of writing, (b) modularized content to accommodate different pedagogies or student needs, (c) opportunities for extended and varied writing practice, and (d) formative writing feedback related to writing proficiency and strategies. In the following sections, we consider how prior technologies have addressed these issues, and then discuss how these design principles have been uniquely implemented within the W-Pal tutoring system.

Automated Essay Scoring and Writing Evaluation

A significant challenge for computer-based writing instruction is the automated assessment of student writing and delivery of meaningful feedback. One advantage is that computer-based tools can evaluate many text features consistently and simultaneously, and apply the same criteria to all essays reliably and objectively. Indeed, automated essay scoring (AES) systems have been developed to facilitate essay grading using statistical modeling, machine learning, natural language processing (NLP), and latent semantic analysis (LSA). Prominent systems include *e-rater* (Attali & Burstein, 2006), *IntelliMetric* (Rudner, Garcia, & Welch, 2006), and *Intelligent Essay Assessor* (IEA; Landauer, Laham, & Foltz, 2003). Overall, AES scoring tends to be accurate. Human and computer-assigned scores correlate around .80 to .85 (Warschauer & Ware, 2006), with 40–60% perfect agreement (exact match of human and computer scores) and 90–100% adjacent agreement (human and computer scores within 1 point; e.g., Attali & Burstein, 2006; Dikli, 2006; Rudner et al., 2006). Over time, AES systems have become embedded within automated writing evaluation (AWE) systems that assign scores along with feedback on errors (e.g., spelling) and may include instructional scaffolds and learning management tools (Grimes & Warschauer, 2010). Examples include *Criterion* (e-rater scoring engine) from the Educational Testing Service (Burstein, Chodorow, & Leacock, 2004), *MyAccess* (IntelliMetric engine) from Vantage Learning (Grimes & Warschauer, 2010), and *WriteToLearn* (IEA engine) from Pearson Education (Landauer, Lochbaum, & Dooley, 2009).

Evaluations of AWE technologies have focused primarily on scoring accuracy, although a few studies have examined instructional efficacy. For example, Shermis, Burstein, and Bliss (2004) examined essay scores for over 1000 high school students, half of whom participated in typical classroom instruction and half of whom used Criterion. The two groups did not differ in holistic essay quality, although Criterion users produced longer essays with fewer mechanical errors. Rock (2007) obtained comparable results in a study with over 1,400 ninth grade students using Criterion. Finally, Kellogg, Whiteford, and Quinlan (2010) experimentally manipulated how much feedback 59 undergraduates received from Criterion on three essays. Students received feedback on all essays, one essay, or none. Holistic essay quality did not differ across conditions, although students who received more feedback displayed fewer mechanical errors in their essay revisions. In sum, Criterion[1] has been successful in improving student essays but primarily for mechanical properties, rather than holistic quality.

Grimes and Warschauer (e.g., Grimes & Warschauer, 2010; Warschauer & Grimes, 2008) have argued for the need to examine users' perceptions of AWE tools in the classroom. Successful deployment of writing technologies may depend upon whether teachers and students view the tools as valid, useful, and usable. Within this framework, Warschauer and Grimes (2008) examined perceptions of Criterion or MyAccess in four schools, obtaining survey and interview data from principals, teachers, and students (sixth to 12th grade). Both systems were perceived to increase students' motivation to write and improve writing quality, but the tools were used infrequently due to curricular conflicts. Students

[1] A literature search did not reveal similar evaluations of other systems.

did not always have time for extra writing assignments and the systems could not support every writing genre that teachers wished to cover. In addition, although the systems seemed to promote essay revising, most revisions focused on mechanics rather than content, organization, or style.

Grimes and Warschauer (2010) later examined MyAccess over a 3-year period in four middle schools. System use was initially infrequent—teachers did not create assignments in the system and students rarely revised. However, use increased over time as teachers became more comfortable with the technology. Survey data revealed both positive attitudes and skepticism. Teachers felt that MyAccess saved time, made teaching easier and more enjoyable, and allowed them to focus on higher level concepts. Teachers also reported that students were more motivated to write. However, teachers doubted the accuracy of the automated scores. They also favored MyAccess for persuasive essay writing but preferred traditional methods for informative, narrative, or analytical genre writing. Teachers also felt that MyAccess was suited to teaching sentence fluency and conventions, but less helpful for covering ideas, organization, voice, and word choice. Similarly, students perceived the system as usable and enjoyable, and felt that it increased their confidence and quantity of writing. However, students had trouble understanding the feedback and felt overwhelmed by the quantity of feedback. Some teachers had to create handouts to help students navigate the "pages of suggestions" from the system. In addition, some students began to focus on improving their scores rather than communicating their ideas.

In sum, research on AWE tools is promising but highlights how efficacy may be hindered by student and teacher perceptions. When users doubt the automated scores or feedback, or find them overwhelming, it is unlikely that the system will achieve its true potential. Another concern may be an emphasis on practice and feedback with less attention paid to strategy instruction or modular design. The fundamental purpose of AWE systems is the facilitation of writing assessment rather than teaching students about writing principles, goals, and strategies. Without such instruction, students may not be prepared to utilize the detailed writing feedback these tools offer. Last, an emphasis on error feedback may not satisfy the principle of formative feedback.

Computer-Based Tutorials for Writing

A few technologies have been created to teach specific writing skills or to scaffold the writing process. For example, the LSA-based *Summary Street* (Caccamise, Franzke, Eckhoff, Kintsch, & Kintsch, 2007; Kintsch et al., 2007) supports students' summarization skills. When students write summaries in the system, they receive graphical feedback showing how well their text captures the source materials. Research with Summary Street has shown that students wrote more effective summaries and spent more time engaged in writing when using the system. Perceptions of the system were also positive: students found the system easy to use and appreciated receiving feedback related to what they needed to fix in their summaries. Similarly, Wolfe et al. (2009) developed a web-based tutor for developing argument, counterarguments, and rebuttals. Evaluations of this system have shown the tutorial instruction improved students' ability to perform these tasks. Overall, such research suggests that computer-based tutorials can be effective for training students on specific strategies related to writing.

Another technology, *Computer Tutor for Writing* (CTW; Rowley & Meyer, 2003) adopted a scaffolding approach in which students wrote essays in an enhanced word processor. The interface provided "workspaces" in which students could view descriptions, examples, and hints related to the writing process, such as goal-setting, drafting, and publishing. A tracking system monitored completion of these tasks. Importantly, CTW did not provide a holistic score for essays, nor were students given error feedback or strategy guidance for improving their essays. Thus, writing support in CTW was instantiated solely as structured guidance during composition. An evaluation of the CTW with 471 middle and high school students (Rowley & Meyer, 2003) revealed no difference between control (i.e., no CTW training, $n = 174$) and experimental conditions (i.e., training with CTW, $n = 298$). Neither group improved from pretest to posttest with regards to essay scores; control participants' scores decreased by about 1%, whereas experimental participants' scores increased by about 2%.

Proske et al. (2012) adopted a similar scaffolding approach with the *escribo* system. In *escribo*, students receive online support for prewriting, drafting, and revising processes, along with feedback about their choices at each stage. Forty-two German university students practiced writing with or without the system in one training session and then wrote an unsupported essay in a posttest session. Overall, students who interacted with *escribo* spent more time planning their essays, which facilitated faster drafting of the text. *escribo* students also spent more time revising their essays and the resulting texts were rated as more comprehensible. Thus, when students are provided with both comprehensive strategy help and informative feedback on their writing process, computer-based tutorials for writing are more effective.

In sum, previous computer-based writing tutors have shown mixed results, which may be attributed to whether feedback was provided. Successful tutors for summarization and argumentation focused on fewer skills but offered feedback on students' performance. The main drawback is potentially their scope; they do not provide comprehensive or modular instruction related to the entire writing process. In contrast, CTW addressed all phases of writing with support for each task, but students did not receive strategy feedback. The system appeared to be of little benefit. However, when structured writing support is combined with feedback, as in *escribo*, empirical evidence suggests that a scaffolding approach can be effective.

The Writing Pal

In the development of W-Pal, we have sought to synthesize key principles of strategy instruction, modularity, extended practice, and formative feedback (McNamara et al., 2011). The interdisciplinary development of the initial version of W-Pal spanned over 3 years with input from cognitive psychology, linguistics, computer science, and English education.

Writing Strategy Modules

The principles of *comprehensive strategy instruction* and *modularized content* were instantiated in W-Pal via nine *Writing Strategy Modules* (see Table 1). The content for these modules were

Table 1
Summary of Strategy Training Module Content and Practice Games

Module	Description of Strategies	Practice Games
Prologue	Introduces W-Pal, the animated characters, and discusses the importance of writing	
Prewriting Phase		
Freewriting	Covers *freewriting* strategies for quickly generating essay ideas, arguments, and evidence prior to writing (*FAST PACE mnemonic*)	Freewrite Feud Freewrite Fill-In
Planning	Covers *outlining* and *graphic organizer* strategies for organizing arguments and evidence in an essay	Mastermind Outline Planning Pump
Drafting Phase		
Introduction Building	Covers strategies for writing introduction paragraph *thesis statements*, *argument previews*, and *attention-grabbing techniques* (*TAG mnemonic*)	Essay Launcher Dungeon Escape Fix It – Introductions
Body Building	Covers strategies for writing *topic sentences* and providing objective *supporting evidence* (*KISS & Tell mnemonic*)	RAM-5 Fix It – Bodies
Conclusion Building	Covers strategies for restating the thesis, summarizing arguments, closing an essay, and maintain reader interest in conclusion paragraphs (*RECAP mnemonic*)	Fix It – Conclusions Dungeon Escape
Revising Phase		
Paraphrasing	Covers strategies for expressing ideas with more *precise and varied wording*, varied *sentence structure*, and *condensing* choppy sentences	Adventurer's Loot Map Conquest
Cohesion Building	Covers strategies for adding cohesive cues to text, such as *connective phrases*, *clarifying undefined referents*, and *threading ideas* throughout the text	CON-Artist Undefined & Mined
Revising	Covers strategies for reviewing an essay for completeness and clarity (*TETRIS mnemonic*), and strategies for how to improve an essay by adding, removing, moving, or substituting ideas (*ARMS mnemonic*)	Speech Writer

developed based on research on writing strategy instruction (e.g., Graham & Perin, 2007) and substantive, iterative input from expert writing educators (Roscoe, Varner, Weston, Crossley, & McNamara, in press). Writing strategies were discussed by three animated agents via lesson videos (15–30 min each). Dr. Julie (teacher agent) explained the strategies, and Mike and Sheila (student agents) demonstrated them (Figure 1). These characters were developed using Media Semantics Character Builder software and text-to-speech voices by Loquendo. For many lessons, multiple strategies were organized by acronymic mnemonic devices, which can facilitate adolescent students' recall and use of writing strategies (e.g., De la Paz & Graham, 2002). Quiz and game-like checkpoints were embedded in the lessons to reinforce the content, and students could take notes. All modules were accessible from a "Lessons Tab" in the W-Pal interface, which allowed users to progress through the modules in a flexible order.

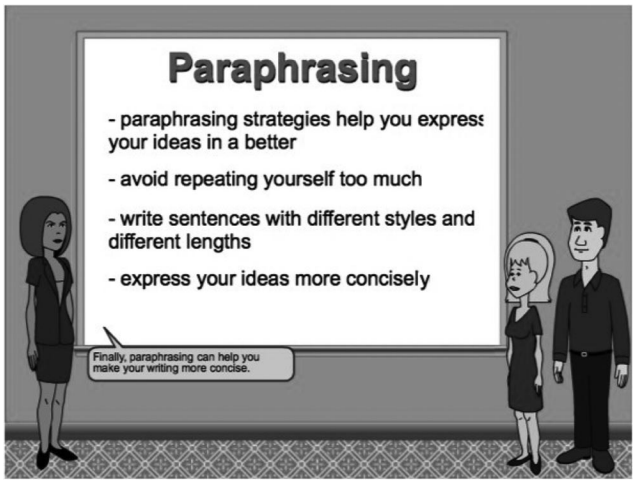

Figure 1. Screenshot of Writing Pal virtual classroom (Paraphrasing lesson).

Game-Based Practice

The principle of *opportunities for extended feedback* was realized by developing two broad modes of practice: *game-based practice* and *essay writing practice*. In W-Pal, a suite of educational games allows students to practice specific strategies outside of the context of complete essays. For example, students can practice strategies for evaluating evidence or building cohesion before applying these strategies in their own persuasive essays. Game-based practice was also chosen to address problems of student engagement. One challenge for ITSs is that students become bored and frustrated with extended practice (Bell & McNamara, 2007; Jackson & McNamara, 2013). Games offer a means of improving students' motivation to participate by leveraging their intrinsic enjoyment of gaming (Shank & Neeman, 2001).

In W-Pal, each Writing Strategy Module was associated with one or more practice games that students "unlock" by completing the lessons (see Table 1). This version offered 15 unique games. These games were iteratively developed by selecting key strategies covered in the lessons and then constructing *generative* or *identification* practice tasks. In generative practice, students write short texts (e.g., a conclusion paragraph) while applying one or more strategies. In identification practice, students examine text excerpts to label the strategies used, or to identify how strategies may be used to improve the text. These practice tasks were then embedded in diverse game mechanics and narratives. Feedback in the practice games was contextualized via the game design, such as winning or losing, earning points, the amount of fuel consumed by a spaceship, or the quality of treasure obtained. Thus, students could judge whether their strategy application was effective based on their

game progress. In some cases, formative feedback was also offered, such as tips for succeeding in the game by using certain strategies or mnemonics.

To provide examples, we briefly present two games: Freewrite Feud and Essay Launcher. In *Freewrite Feud* (see Figure 2), students were given several minutes to freewrite on a persuasive writing prompt. For each prompt, a hidden list of key words and concepts was constructed based on a previous corpus of freewrites. Students earned points by typing their ideas quickly and continuously, and earned additional points when their freewrites incorporated up to six of the key words. Because these key words were hidden from the player, this generative game encouraged students to practice brainstorming many ideas, arguments, and potential pieces of evidence because doing so would trigger the key words and earn a higher score.

Essay Launcher was an Introduction Building game (see Figure 3). In this identification game, students attempted to repair and rescue several spaceships. To "repair the ship," students chose a thesis statement for an example introduction paragraph from a list of three options. To "set the course," students turned a dial labeled with attention-grabbing techniques to identify the technique used in the paragraph. Once both selections were made, students consumed one fuel unit to launch the ship. If either choice was incorrect, the launch malfunctioned. Students then received feedback about introduction strategies and could try again. Points were based on rescued ships and remaining fuel. This game allowed students to practice evaluating key characteristics of essay introductions.

Essay-Based Practice and Feedback

The principles of *formative feedback* and *opportunities for extended practice* were supported by the W-Pal *Essay Writing Interface* (see Figure 4). W-Pal allowed students to practice writing timed persuasive essays using SAT-style prompts in which they could synthesize and apply strategies covered in any module. Students could select the prompt, set the time limit, and use a scratchpad for prewriting. Essays were written using a simple word processor and then submitted for automated assessment.

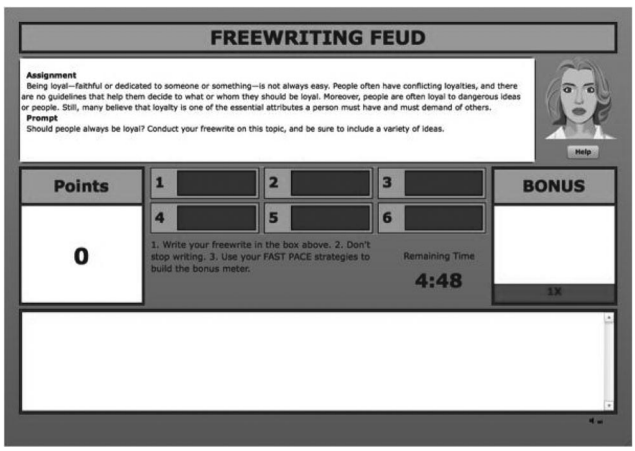

Figure 2. Freewriting Feud practice game (Freewriting).

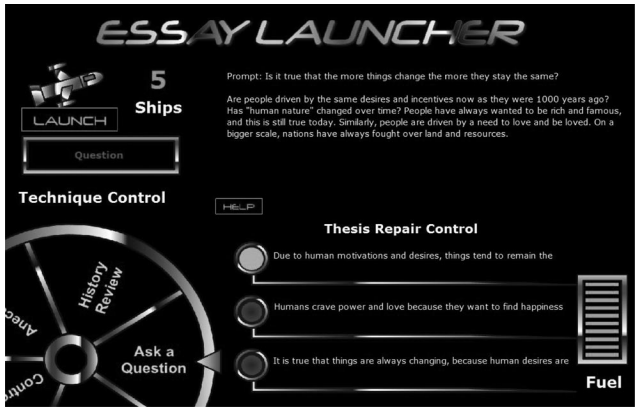

Figure 3. Essay Launcher practice game (Introduction Building).

W-Pal scoring is powered by NLP algorithms utilizing Coh-Metrix and other text analysis tools (Crossley & McNamara, 2011; Graesser & McNamara, 2012; McNamara, Crossley, & McCarthy, 2010; McNamara, Crossley, & Roscoe, 2012), and such algorithms are a key source of the *intelligence* of a writing ITS. Within ITSs that accept natural language as input (e.g., essays or verbal explanations of scientific processes), 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 (McNamara, Crossley, & Roscoe, in press). In service to these goals, W-Pal utilizes Coh-Metrix to analyze text on several dimensions of cohesion including co-referential cohesion, causal cohesion, density of connectives, lexical diversity, temporal cohesion, spatial cohesion, and LSA. Coh-Metrix also calculates syntactic complexity and provides psycholinguistic data about words (parts-of-speech, frequency, concreteness, imagability, meaningfulness, familiarity, polysemy, and hypernymy).

Essays submitted to W-Pal initially received a holistic rating from *poor* to *great* (6-point scale). Writers also received feedback that addressed particular writing goals and strategy-based solutions (see Figure 5). Such feedback was implemented as a series of scaffolded, threshold-based algorithms based on different linguistic properties and categories: *legitimacy* (e.g., proportion of nonwords), *length* (e.g., number of words), *relevance* (e.g., occurrence of key words), and *structure* (e.g., number of paragraphs). For example, writers whose essays lacked elaboration (i.e., short essays) might receive feedback such as, "One way to expand your essay is to add additional relevant examples and evidence," and prompts such as, "Have you created a flow chart or writing road map to help you organize your ideas?" The feedback also directed students toward relevant lessons or practice games. Importantly, feedback scaffolding helped to deliver only the most appropriate help; feedback was delivered only for the lowest threshold failed in the series of checks. We assumed that students who struggled to produce *any* text may not be ready to implement feedback about cohesion. Instead, these students may gain more from planning. If essays passed basic thresholds, they received feedback encouraging overall revision. Depending on the quality of individual sections, essays also received formative feedback for introduction,

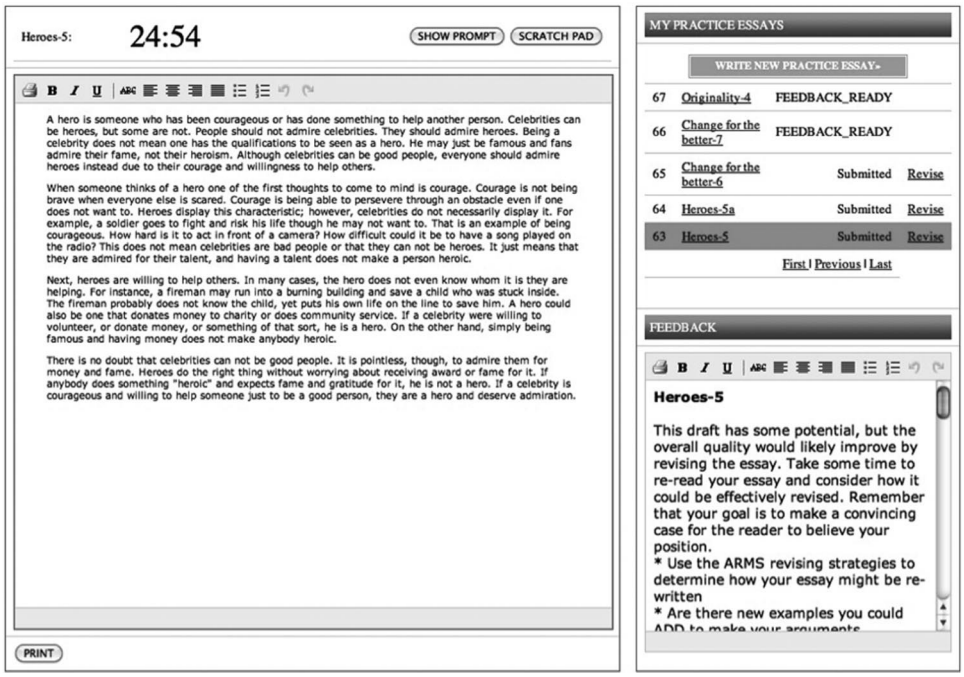

Figure 4. Essay Writing Interface.

body, and/or conclusion building strategies. For instance, an essay lacking a clear body might receive feedback stating "good writers often review their writing flowchart or an outline. Think about the best order and organization of the body paragraphs," and asking, "Could a stranger understand your ideas without further explanation?" (Figure 5).

Unlike previous AWE systems, W-Pal focuses on strategy instruction and formative feedback and provides no specific error feedback on style, mechanics, spelling, or grammar. Spelling and grammar errors are relatively easy to detect, but assessing the quality and relevance of thesis statements, topic sentences, examples, counterarguments, and many other essay elements is more difficult. In the case of thesis statements, for example, it is a nontrivial matter to determine which sentence writers intended to communicate their position, if any. Once this determination is made, one must assess how the thesis relates to the prompt, subsequent arguments, and argument structure. At this stage of W-Pal development, we focused on the broader categories, which necessarily limited the specificity of W-Pal feedback.

In sum, development of W-Pal has sought to satisfy four central design principles that emerge from the ill-defined nature of writing, which has not been demonstrated in previous technologies for writing instruction. A fundamental question for deployment was whether an intelligent tutor for writing could be feasibly implemented with our target population of high school students. Would students use the system? Would students perceive a "computer tutor" as a viable instructional resource? To address these questions, we conducted a feasibility study in five high school English classrooms throughout a school year. Because our primary purpose was to assess feasibility, we did not employ a controlled experimental design (i.e., comparison to non-W-Pal instruction) or ablative design (i.e., selective removal of system features). Thus, strong conclusions about efficacy cannot be drawn about the impact of W-Pal from this study.

Method

Participants

The intended users of Writing Pal are English-speaking high school students. Two high school English teachers and 141 10th grade students participated in this study over 6 months (November, 2010 to May, 2011) with their English classrooms. Teachers were asked to use the entire W-Pal, including Writing Strategy Modules, practice games, and essays. However, they were not given strict rules for how W-Pal was to be integrated (e.g., module order, assignment pacing and duration, or curriculum integration). Teachers and students (via their teachers) could contact the W-Pal team for technical support and teachers had weekly conference calls with the researchers. The participating high school was located in the Washington, DC area, and enrolled over 2,400 students. The school enrolled 49.0% female students, with 22.3% Asian, 4.2% Black, 9.0% Hispanic, and 59.9% White students; 7.0% of students were described as limited English proficiency, and 10.9% qualified for free or reduced-price meals.

Measures

Data logging. As students interacted with W-Pal, their access of system tools was logged. To examine usage of W-Pal, we considered access and completion of the lesson videos, frequency of games played, and frequency of essay submissions.

Lesson perception survey. After viewing each lesson, a five-item survey appeared. Using 4-point scales, students rated "how

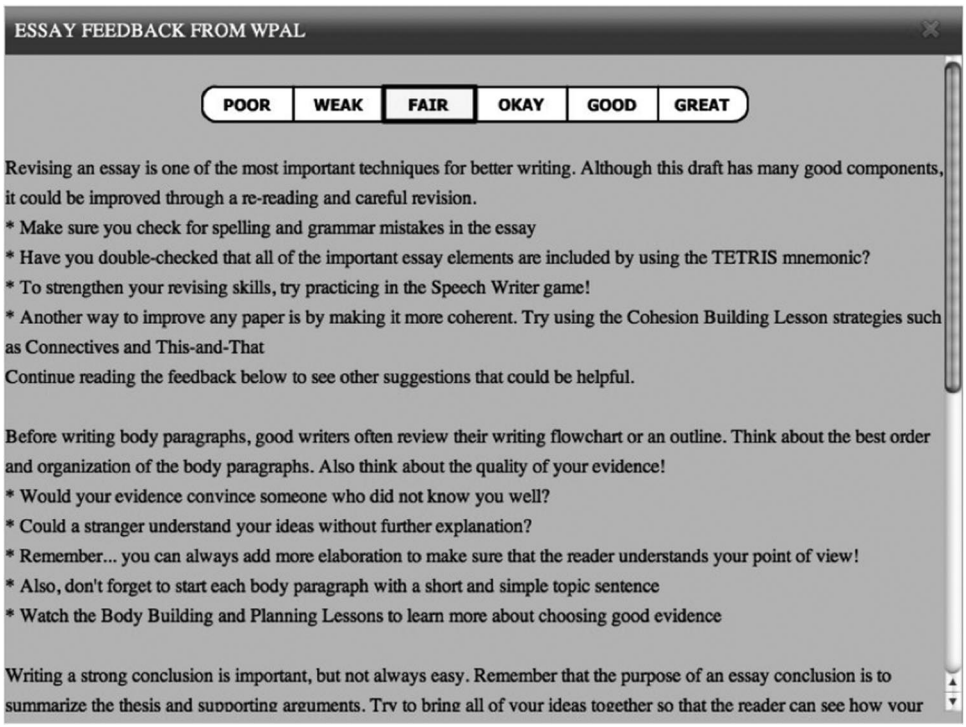

Figure 5. Example essay feedback report. WPAL = Writing Pal.

many new ideas" they learned (i.e., 0, 1–2, 3–4, or 5 or more ideas) and whether they would be willing view the lesson again. In open-ended items, students were asked to describe the "most helpful information" they learned, describe their perceptions of the animated characters, and provide suggestions for "how to improve this lesson."

Game perception survey. After interacting with W-Pal for several months (i.e., in February), students were asked to complete a four-item feedback survey of their perceptions of the games. Using 4-point scales, students rated a sampling of 11 games regarding helpfulness for practicing writing strategies, and rated the games regarding enjoyment. In two open-ended items, students were asked to provide suggestions for improving the helpfulness of the games and redesigning the games to be more enjoyable and engaging.

Feedback perception survey. In addition to the Game Perception Survey, students completed an eight-item survey of their perceptions of the essay writing tools and feedback. Using 4-point scales, students rated the overall difficulty of using the essay writing interface, the difficulty of specific tools, feedback quantity, understandability of the feedback, and usability of the feedback. In two open-ended items, students were asked to offer suggestions for making the feedback "more clear, more understandable, or more usable" and to suggest what "essay features or writing strategies" should be included in future feedback.

Pre- and post-study essays. Students wrote timed (25 min), prompt-based essays on two SAT–style prompts regarding "competition" and the influence of "images and impressions." These essays were written offline (i.e., not within W-Pal), manually transcribed by the research team, and scored via natural language algorithms powered by Coh-Metrix (Crossley, Roscoe, Graesser, & McNamara, 2011). The accuracy of this algorithm, based on a separate test set of 105 essays and expert human scores, was 39% perfect agreement and 92% adjacent agreement. Descriptive information was also calculated for each essay, including the number of words, sentences, paragraphs, and sentences per paragraph. Text cohesion was assessed in terms of argument overlap (i.e., average overlap between head nouns and pronouns in adjacent sentences), given/new information (i.e., a Latent Semantic Analysis score indicating the amount of given compared to new information), and lexical diversity (i.e., degree to which a variety of words versus the same words are used across the text, using the measure D; Malvern, Richards, Chipere, & Durán, 2004). Prior research has indicated that higher quality essays are associated with a decrease in cohesion and an increase in lexical diversity (Crossley & McNamara, 2011). We also examined measures of lexical sophistication typically associated with essay quality (e.g., Crossley, Weston, McLain Sullivan, & McNamara, 2011), including word concreteness, word hypernymy (i.e., specificity), and the number of hedging words (i.e., an indicator of uncertainty).

Procedures

Students wrote a pre-study essay in November. Throughout the school year, teachers incorporated W-Pal into their English classroom curriculum. Students viewed the lessons, played the games, wrote practice essays, wrote essays assigned by teachers, and completed the surveys. Essays assigned by the teachers

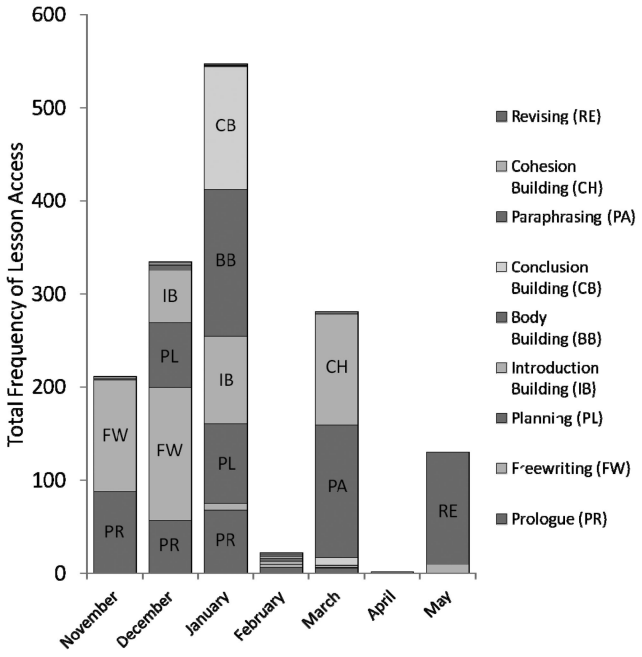

Figure 6. Total frequency of lesson viewing across a 6-month time period.

often explicitly linked to reading assignments, such as Moliere's *Tartuffe*. Students wrote a post-study essay in June. As this was an ecological setting, some students did not complete all assignments.

Results

Students' Use of the System

Students interacted with W-Pal for about 16 total hours, on average, but students' use of W-Pal was unevenly distributed by module and across time. Figure 6 shows the distribution of strategy lessons accessed over the 6 months of the study (substantive activities are labeled with an abbreviation of the module name). Access was defined as a student interacting with at least one complete segment of the lesson. One pattern is that teachers mainly followed the sequence of prewriting, drafting, and revising. That is, they assigned the modules linearly in the "order" they were listed in the W-Pal interface. Teacher interviews indicated that they discouraged exploration; they preferred students to focus on current assignments and not to "get ahead." Second, most use of W-Pal lessons occurred during the first 3 months and then became more sporadic. January was particularly active as teachers encouraged students to complete the prewriting and drafting modules in preparation for SAT practice tests. Teachers did not assign lessons during February and April. Teacher interviews indicated that these months were devoted to separate writing assignments (e.g., a "how-to" paper), literature instruction (e.g., *Tale of Two Cities* and *Things Fall Apart*), and preparation for state exams.

Over time, lesson activity appeared to decrease. This pattern is substantiated by the average completion percentage of each module (Table 2). In general, students seemed more likely to complete the earlier modules (e.g., Freewriting), but tapered off in the later modules (e.g., Revising). One explanation may be student fatigue. After 5 months of using W-Pal, any novelty had likely diminished. In addition, teachers' focus on literature assignments and test

Table 2
Average Completion Percentage for Lesson Videos, Frequency of Game Play, and Maximum Number of Game Plays by Module

	Lesson completion		Game play		
Module and game	*M*	*SD*	*M*	*SD*	Maximum
Prologue	86.0	27.5			
Freewriting	90.2	28.7			
Freewrite Feud			0.59	0.89	4
Freewrite Fill-In			0.59	0.85	3
Planning	83.4	35.9			
Mastermind Outline			0.72	0.98	6
Planning Pump			0.76	0.82	4
Introduction Building	82.9	31.7			
Dungeon Escape			0.86	0.88	4
Essay Launcher			0.45	0.60	4
Fix-It			0.49	0.61	3
Body Building	82.9	37.2			
RAM-5			0.15	0.36	1
Fix-It			0.29	0.45	1
Conclusion Building	73.1	44.1			
Dungeon Escape			0.53	0.77	4
Fix-It			0.36	0.51	2
Paraphrasing	78.2	40.7			
Adventurer's Loot			0.44	0.51	2
Map Conquest			0.53	0.68	5
Cohesion Building	54.3	49.3			
CON-Artist			0.31	0.56	4
Undefined & Mined			0.54	1.14	6
Revising	68.5	45.1			
Speech Writer			0.29	0.54	3

preparation may have led to a decreased emphasis of W-Pal in the classroom.

Figure 7 provides a similar visualization of students' game playing across modules, with substantive activity labeled by module. Few games were played in November, as most students had not unlocked any games. However, more games were played in the following months once teachers assigned the planning and drafting modules. Interestingly, game play continued during February when no new modules were assigned. Interviews revealed that teachers encouraged students to use the games as further practice during this time. In the final months, however, students mainly accessed the games associated with assigned modules. Table 2 shows the mean frequency of playing each game. Games encountered earlier in instruction (e.g., Mastermind Outline), were played slightly more often than later games (e.g., Speech Writer). However, there was variation in game play and some games from later modules were played as often as earlier games. The overall low frequency of play is likely a result of teachers' discouragement of exploration. The wide variety of games offered by W-Pal may have also contributed. With many games to choose from, the desire to "master" any one game might have been low.

Use of the essay writing tools was somewhat sparse because teachers used W-Pal for specific assignments rather than self-selected practice. Teachers assigned two to three W-Pal practice essays with automated feedback (on "Honesty," "Uniformity," or "Heroes") in December and January. Students were not required to revise these essays and course grades were based only on assignment completion. In April and May, teachers assigned students to write on the "Memories" prompt in relation to the novel *Things Fall Apart* (with automated feedback). Revising of this essay occurred outside of W-Pal via extensive peer reviewing. Teachers initially reported confusion about how teacher-created prompts differed from built-in W-Pal writing prompts—essays written on teacher-created prompts could not be assessed by the algorithm in this version of the system. However, after discussion about this functionality, teachers still chose to create two new assignments in W-Pal. In one essay, students wrote about interpersonal perceptions in relation to the novel *Tartuffe* (January), and students responded to a newspaper article about the value of study halls in high schools (February).

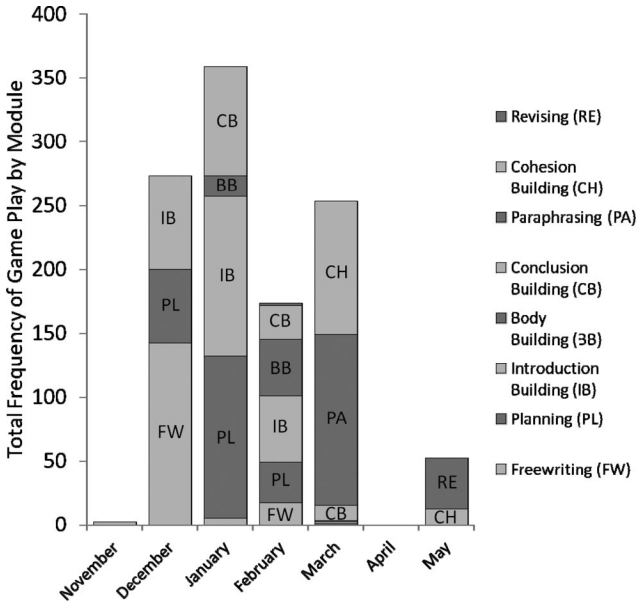

Figure 7. Total frequency of game playing across a 6-month time period.

Interviews revealed that teachers perceived W-Pal's essay tools favorably, and felt that the system allowed them to assign more writing that was feasible without W-Pal. Specifically, W-Pal provided an accessible means for students to practice writing, with automated feedback, and teachers could access these essays and feedback online. W-Pal also provided several ready-made writing prompts for assignments. However, the system could not support the full range of writing assignments that were required in the curriculum, a common problem for AWE systems (e.g., Grimes & Warschauer, 2008). Different writing genres (e.g., journalism and narrative) possess unique constraints that cannot be assessed by the same algorithm; computational linguistics models must tailored to each type. Most systems, including W-Pal, have focused upon persuasive writing due to its importance for standardized testing. Other genres are not currently supported but are a target for future development. Teachers also understood that W-Pal was still "in development" and thus were somewhat wary of basing students' grades on W-Pal assessments. This concern may also have limited the number of practice essays teachers assigned. Teachers may have been hesitant to utilize W-Pal for writing practice unless they could also review or grade the assignments independently. Teachers understood the scoring and feedback procedures but, as conscientious instructors, they wanted to remain actively aware of and involved with their students' work and progress.

In sum, students used a variety of W-Pal features but did so unevenly over the year. W-Pal deployment was not a smooth and continuous process; as with any educational resource, teachers were selective and opportunistic about how and when to use the system. Results also suggest that engagement with the system declined over time. We next consider students' perceptions of W-Pal and how such perceptions may have impacted system use and feasibility.

Lesson Perceptions

Figure 8 (left side) presents the percentage of students as a function of the number of ideas they reported having learned from the lessons. In general, students reported the lessons to be helpful and informative. On average and across lessons, over half of the students (55.8%) reported learning three or more ideas per lesson. Within the open-ended questions asking students to summarize *the most helpful idea* learned from the lessons, the mnemonic devices were the most frequent response. Thus, students seemed to value and remember the acronyms such as *TAG*, *RECAP*, and *ARMS* designed to cue recall of specific strategies. In contrast, students disliked the presentation of the lessons (Figure 8, right side). On average and across lessons, many students viewed the characters as *awkward* (62.3%) and *boring* (60.6%), but still *informative* (30.4%).

In open-ended responses (see Table 3), students critiqued agent dialog and requested succinct instruction with more competent and less "cartoonish" characters. The computerized voices were also unpopular, in part because of a text-to-speech glitch that sometimes caused overlapping speech. Both students and teachers re-

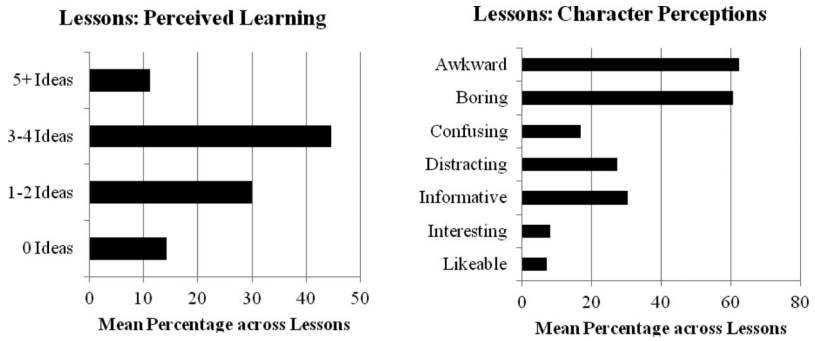

Figure 8. Student perceptions of learning and animated characters in Writing Pal lessons.

quested that the lessons be shorter and faster, while retaining all of the information. These concerns are summarized by one student, who commented, "the little jokes between characters aren't amusing, especially in the monotone computer voices. Cutting out all the unnecessary dialogue between characters would shave off a good amount of time." Altogether, these results support the earlier hypothesis that lesson use decreased due to fatigue. As students progressed through the lessons, and encountered the same design issues, students' willingness to engage with the lessons likely decreased.

Game Perceptions

Across sampled games, students (n = 116) reported the games to be *somewhat helpful* (50.5%) or *very helpful* (29.6%) for practicing the writing strategies (Figure 9, left). Similarly, students reported the games to be *somewhat enjoyable* (46.4%) or *very enjoyable* (19.1%) to play (Figure 9, right). Thus, most students felt that the games they played were beneficial and generally engaging. Open-ended comments (see Table 3) highlighted ways in which students felt the games could be improved. For example, one student requested that we make the games "*more challenging* [because even] if I hadn't taken the W-Pal lessons, I would have been able to complete the challenges with fairly high scores." Other students expressed interest in further generative practice, such as "when we learn the strategies, I think should be a challenge where we actually use the strategy instead of finding them in essays." Another student suggested that "the games could be more difficult and more

Table 3
Student Responses and Recommendations Regarding Strategy Lessons and Practice Games

Observation	Examples
1. Students valued the strategies and mnemonics.	"FAST PACE is going to help me write better essays! I learned important acronyms, and information. I learned to think about the prompt, add questions, think about the opposing side" "The TAG mnemonic and the attention grabbing techniques were very helpful for making me understand introductions better" "RECAP—restate, explain ideas, closing, avoid new things, present interestingly"
2. Students disliked the length and presentation style of the lessons.	"Their voices are very robotic and the lesson was way too long, maybe if it was split into several sections then it would be easier to concentrate on the task" "Had very good information but I disliked the synthesized voices" "The information is good but I lost interest throughout the lesson. I feel like I would learn a lot more if the information went faster and was straightforward"
3. Students desired games that were more difficult and interactive.	"When we learned the strategies, I think there should be a challenge where we actually use the strategy instead of finding them in essays" "Make the challenges more challenging. Even if I hadn't taken the W-Pal lessons, I would have been able to complete the challenges with fairly high scores"
4. Some students found the game instructions inadequate.	"Some of the instructions were hard to follow" "I had a little trouble understand exactly what to do with the directions."
5. Students suggested improvements in the game graphics and sound.	"The games could have better graphics and music to make the games more enjoyable" "The games are slow and the graphics are not the best, so unfortunately, the games become boring which weakens their effectiveness"
6. Students requested that more game elements be added.	"Many of the games were not very fun because they had a learning element that was very obvious. It would be better if the element was not as obvious, so the game was more fun. Basically, more pictures and music and less words" "Make it a point system and make it a competition amongst our peers"

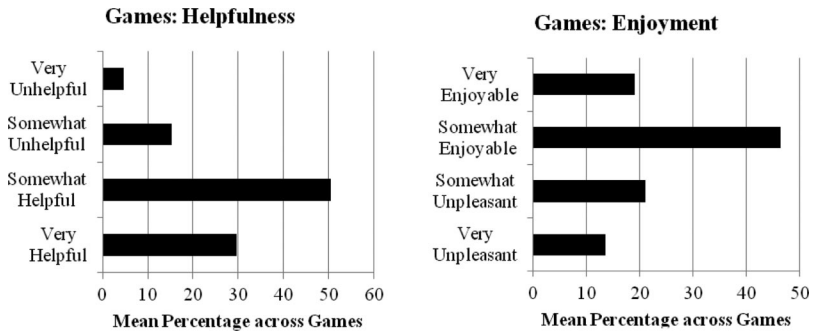

Figure 9. Student perceptions of the helpfulness and enjoyment of games.

interactive for learning these writing strategies rather than just reading." In sum, students valued the games, but positive perceptions may have been impacted by games that lacked challenge, opportunities for interaction, or clear directions.

Essay Writing and Feedback Perceptions

Overall, students (n = 103) rated the essay writing tools as *easy* or *very easy* (81.5%) to use (Figure 10, top left). However, two features frustrated some students: 23.7% of students reported that reading the feedback was *somewhat* or *very difficult*, and 24.6% felt that revising their essays was *somewhat* or *very difficult*. This may have been due to feedback quantity or clarity (Figure 10, top right). Although most students reported that they received *just the right amount of feedback* (49.5%), others reported that they received *not enough* (38.8%) or *too much* (11.6%). From internal testing (Roscoe, Varner, Cai, Weston, Crossley, & McNamara, 2011), we knew that feedback quantity could be variable. Essays that failed a basic check (e.g., length) received only one feedback message. However, essays that advanced further could receive more messages on multiple topics. These extremes may have led to perceptions of insufficient or overwhelming feedback, respectively. Similarly, as shown in Figure 10 (bottom left), most students rated the feedback as *understandable* (61.2%), but some students rated the feedback as *somewhat confusing* (29.1%) or *very confusing* (9.71%). Despite these challenges, students rated the feedback as useful (Figure 10, bottom right) *occasionally* (45.6%) or *often* (33.0%).

Students' open-ended responses (see Table 4) further highlighted student concerns. Specificity was a particular critique;

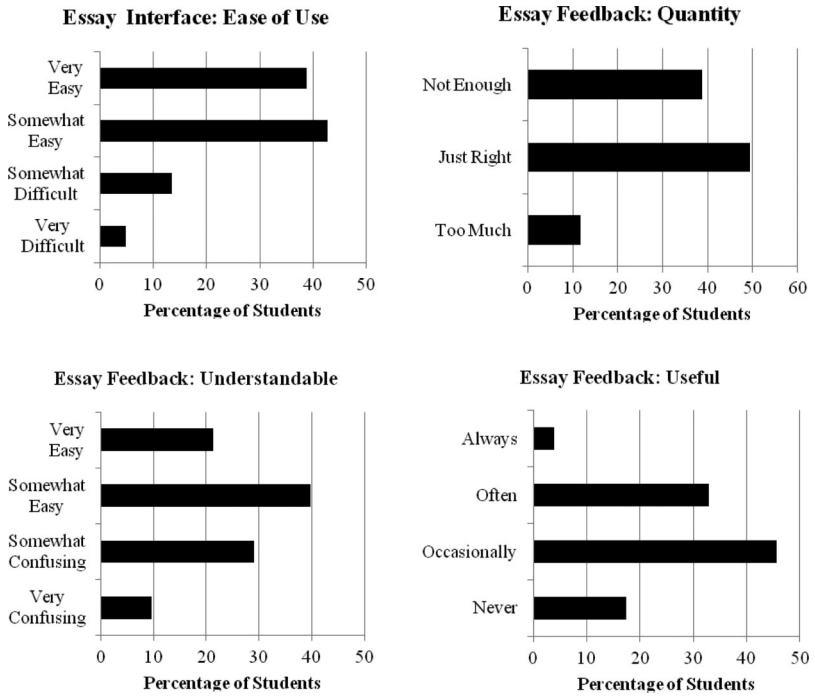

Figure 10. Student perceptions of ease of use, quantity, understandability, and usefulness of automated essay feedback.

Table 4
Student Responses and Recommendations Regarding Essay Scoring and Feedback

Observation	Examples
7. Students requested more specific feedback.	"The feedback should show what specific things made me get the grade" "The feedback needs to be more helpful for us on our own personal essay. Not just general feedback. I don't know what I did wrong in my essay when you just give a general understanding of it"
8. Students requested more individualized feedback.	"My introduction and my supports. I still have a hard time finding supports that directly answer the question" "I would like to know in the feedback if my examples were not strong enough, if I had a weak thesis, things like that"
9. Students expressed conflicting concerns about the quantity of feedback.	"Use less feedback and cut straight to the point of what the essay needs and give examples" "The feedback is very brief. W-Pal never really tells you what you need to improve on."
10. Some students expressed skepticism at the speed and accuracy of scoring.	"I do not like how the essay is graded in less than a second! I feel my essay is not being graded properly and I don't feel I have been given accurate feedback" "You cannot grade an essay in 5 seconds! Everybody gets the same grading of "fair." I can't use it if I don't believe that it is true."

students commented that "the constructive criticism could be a bit more detailed on what the writer needs to work on instead of an overview" or should "be specific and give us exact examples on what we should do to improve our writing." Other students requested that the feedback system provide information on both the strengths and weaknesses in an essay, e.g., "the automatic feedback could also give you good points on your essay, what was strong and what you should continue do." Thus, the provision of feedback at the level of broad categories (e.g., body building strategies) rather than specific essay elements (e.g., evidence quality) was helpful but inadequate for some students. Overall, the feedback provided by W-Pal in this study was perceived as beneficial and relevant to students' needs, but the content of the feedback should be expanded to address more detailed issues.

Essay Quality

The natural language algorithm analyses of pre-study and post-study essays ($n = 113$) are provided in Table 5. Essay scores increased significantly from a mean of 2.3 ($SD = 0.8$) prior to the study to a mean of 2.9 ($SD = 0.8$) after the study, $t(112) = 5.85$, $p < .001$, $d = 0.71$. Associated with these gains were positive changes in essay structure and lexical sophistication (see Table 5). Post-study essays were longer, containing more words and sentences. Essays also showed a clearer paragraph structure, with more paragraphs overall and somewhat fewer sentences per paragraph (e.g., fewer students wrote one-paragraph essays). Post-study essays improved in vocabulary use, including more concrete wording, more precise wording (word hypernymy), fewer hedging words (e.g., *maybe* or *might*), and greater diversity. Finally, essays showed more developed and elaborated content with less repetition of themes (less overlap of arguments and given information) and wording (increased lexical diversity).

Given the patterns of W-Pal use throughout the feasibility study, it would be unlikely to observe strong effects of using the system on essay gains. W-Pal was only one component of a broader curriculum. Nonetheless, to assess how and whether use of W-Pal

Table 5
Essay Characteristics for Pre- and Post-Study Timed Essays

Measure	*M* (*SD*)		*t*(112)	*p*
	Pre	Post		
Essay score	2.30 (0.84)	2.88 (0.79)	5.85	<.001
Length				
Number of words	260.81 (76.38)	308.27 (84.49)	6.49	<.001
Number of sentences	15.46 (5.10)	18.27 (5.13)	5.66	<.001
Structure				
Number of paragraphs	3.43 (1.32)	3.97 (0.83)	3.87	<.001
Sentences per paragraph	5.33 (3.02)	4.72 (1.44)	−1.83	.071
Cohesion[a]				
Argument overlap	0.51 (0.17)	0.41 (0.14)	−5.04	<.001
Given/new information	0.32 (0.04)	0.30 (0.04)	−4.43	<.001
Lexical diversity	85.13 (21.39)	98.29 (21.56)	5.27	<.001
Lexical sophistication				
Word concreteness	387.00 (32.29)	405.53 (30.78)	3.87	<.001
Word hypernymy	1.57 (0.23)	1.66 (0.19)	4.23	<.001
Hedging words	14.2 (10.6)	9.9 (7.6)	−4.10	<.001

[a]These cohesion indices indicate the extent to which arguments, ideas, and words are repeated across sentences and throughout the text.

might have influenced writing proficiency, an exploratory linear regression analysis was conducted to identify potential predictors of post-study essay quality. Eight predictor variables were simultaneously entered. As measures of students' prior writing ability and knowledge, *pre-study essay scores* and self-reported *grade-point average* (GPA) were included. As indicators of system use, we included students' percentage completion of *prewriting lessons* (Freewriting and Planning), *drafting lessons* (Introduction Building, Body Building, and Conclusion Building), and *revising lessons* (Paraphrasing, Cohesion Building, and overall Revising). Similarly, we included the frequency of game play within each phase: *prewriting games* (Freewrite Feud, Freewrite Fill-In, Mastermind Outline, and Planning Pump), *drafting games* (Essay Launcher, Dungeon Escape, Fix It, and RAM-5), and *revising games* (Adventurer's Loot, Map Conquest, Undefined & Mined, CON-Artist, and Speech Writer). Because teachers chose to restrict essay writing practice, there was little variability in essay writing, and this variable was not included.

The resulting linear regression model was significant, $F(112) = 2.93$, $p = .005$, $R^2 = .18$, accounting for about one fifth of the variance in post-study essay scores (see Table 6). Two variables were predictive of essay quality: *pre-study essay scores* and viewing of the *drafting lessons*. Interestingly, students' prior writing ability (pre-study essay score), but not their GPA, was a significant predictor of post-study essay quality. These results suggest that writing skill was not solely a function of students' prior academic abilities, but reflected knowledge of specialized skills and strategies related to writing. Students' completion of the drafting lessons was positively associated with their writing development above and beyond prior writing ability. Drafting lessons are perhaps the most immediately relevant to students' writing of timed essays, because they provide direct strategies for generating essay text. Overall, although we cannot conclude that W-Pal directly improved students' writing, these results tentatively support the feasibility of intelligent tutoring of writing in high school classrooms.

Discussion

The unique design of W-Pal was informed by the ill-defined nature of writing, in which there is significant ambiguity and subjectivity with respect to pedagogy and assessment. We have sought to provide comprehensive and modular strategy instruction, diverse opportunities for extended practice, and formative feedback on students' writing. In this study, we evaluated how W-Pal was perceived by high school in English classrooms. A fundamental assumption was that feasibility depends on whether users view the system as a valid and valuable tool for instruction and feedback. Thus, students' use and perceptions of W-Pal were the central focus.

Our results suggest that this initial version of W-Pal was generally well received. Most components of W-Pal were judged as beneficial sources of writing instruction, practice, and feedback. Students could describe specific content that they learned from the lessons and games, and rated these tools and essay feedback as helpful and easy to use. Students seemed to view a "computer tutor" as a worthwhile addition to the English classroom curriculum. Preliminary evidence also suggests that students benefitted from using certain W-Pal tools. Thus, the initial iteration of W-Pal was feasible with regards to positive user perceptions and usage.

Our results also highlighted several problems to overcome that may undermine long-term feasibility and potential efficacy. First, students felt that the lessons were too long and didactic, and disliked the cartoonish characters in the lessons. In some ways, the lengthy lesson videos were too similar to a *presentational* mode of writing instruction described by Hillocks (1984). Hillocks contrasted writing outcomes for interventions that employed different instructional modes and content. The most effective instruction occurred in an *environmental* mode wherein instructors minimized lecturing and focused on specific objectives and strategies, with ample opportunities for scaffolded practice. In contrast, instruction was less effective in the prescriptive and teacher-dominated presentational mode. Although interactive checkpoints were included in the lessons, students' overall perceptions were that the lessons were too long, boring, and lecture-like. This lesson structure may also have insufficiently met the goal of providing modular instruction; each lesson video comprised multiple strategies related to multiple goals. A series of shorter lessons, each with a focus on one or two related strategies, may have been more germane to Hillocks' environmental mode. Students could iterate between lessons and practice more flexibly, and instructors could be more selective with the content they wished to cover.

More broadly, the issue of information density within instructional modules speaks to the appropriate grain-size of ITS instruction in ill-defined domains. When learners must make many strategic decisions to enact a task, instruction may need to focus initially on fewer decisions before asking students to synthesize them. With each additional, simultaneous strategy choice, it becomes more difficult for learners to perceive the impact or utility of each strategy. In problem-solving domains (e.g., physics), re-

Table 6
Linear Regression Analysis to Predict Post-Study Essay Scores

Variable	*r*	*B*	β	*SE*	*t*	*p*
Pre-study essay score	**.31**	**0.273**	**.293**	**.088**	**3.09**	**.003**
GPA	.12	−0.031	−.026	.118	−0.26	.794
Prewriting lessons	.05	−0.002	−.131	.002	−1.02	.308
Drafting lessons	**.17**	**0.004**	**.431**	**.001**	**2.81**	**.006**
Revising lessons	.08	−0.001	−.157	.001	−1.20	.233
Prewriting games	−.07	−0.015	−.053	.032	−0.47	.640
Drafting games	.00	−0.054	−.191	.035	−1.54	.126
Revising games	.12	0.058	.181	.037	1.59	.115

Note. GPA = grade-point average. Estimated constant term is 2.32. Boldface font indicates statistically significant predictors.

search has shown benefits for systems that require students to specify each step of their solution process rather than merely the final answer (Hausmann, VanLehn, Nokes, & Gershman, 2009; VanLehn et al., 2005). This decomposition allows the system to assess and provide feedback for individual steps, and learners are encouraged to consider the impact of each decision. Analogously, in intelligent tutoring in ill-defined domains such as writing, it may be beneficial to teach fewer writing strategies at one time so that students can more gradually build up to the full complexity of the writing process. The modular content of the ITS should facilitate the decomposition of complex processes into manageable units for initial learning, which can be subsequently recombined and applied strategically in later practice.

A second critique expressed by students related to types and difficulty of learning tasks presented in the educational games. Surprisingly, some students expressed interest in more difficult games that required active generation of text. Such students wanted to practice by applying strategies to their own writing rather than inspecting examples written by others. Not surprisingly, we also observed a high degree of variability in students' game preferences. Games that were played frequently or rated highly by some students were despised by others, and vice versa. Only a few games were broadly disliked; for instance, *RAM-5* (a body building game in which students matched potential evidence to topic sentences) had little replay value, and the task was vague. A few games were liked by the majority of students. One example was *Map Conquest*, a Risk-like game in which students earn resources by identifying paraphrasing strategies and then use those resources to "conquer" a map controlled by computer opponents. An interesting facet of this game is that the learning task (identifying paraphrases) and the game task (taking over the map) are disjoint. Success in the learning task did not guarantee success in the game, and vice versa. This might have made the "gaming" aspects of the practice more salient for some students.

The positive perception of educational games in W-Pal suggests that this could may a valuable component for intelligent tutoring in ill-defined domains. Specifically, games may help to offset some of the motivational threats that undermine students' engagement with ITSs and extended practice. Success in ill-defined domains requires learning of underspecified concepts and relations, and the ability to recharacterize problems to apply available strategies (Lynch et al., 2009). Developing such skills may be frustrating as students struggle to master many decisions and tasks. Indeed, students often report high apprehension and low confidence regarding their writing abilities (e.g., Pajares, 2003). Our results hint that educational games may help to ameliorate some of the affective challenges that arise with learning in ITSs and ill-defined domains (e.g., Craig, Graesser, Sullins, & Gholson, 2004). Games may provide a more pleasant setting where practice is embedded within an enjoyable experience, and feedback is framed within game mechanics or narrative rather than overt critique. However, based on these findings, developers who wish to bolster ITSs with educational games should ensure that the games offer sufficient challenge, promote generative activity, and exhibit varied gameplay.

A final concern revealed by the study, and perhaps the greatest challenge for future development, was the need for more specific and individualized feedback. Students expressed a clear desire to learn more about the individual strengths and weaknesses of their essays, and a lack of such specificity undermined confidence in the system for some students. However, improvements to W-Pal's feedback engine will require sophisticated additions and refinements to underlying computational linguistics algorithms. Although the framing and content of the feedback is paramount—feedback must be well-constructed to provide actionable suggestions in a scaffolded and nonthreatening manner—the feedback process is necessarily constrained to essay features that can be reliably detected. We are currently exploring alternative methods for developing feedback algorithms.

Issues of valid and formative feedback generalize beyond essays and writing. Algorithm development is likely to be a key obstacle in the growth of tutors for writing and other ill-defined domains (McNamara et al., in press). Any ITS that accepts open-ended or natural language input, and attempts respond to learners with intelligent guidance and help, may need to solve a similar set of problems. For example, an ITS that allows users to explain scientific concepts will require algorithms that can process and interpret users' intended answers. Tutorial feedback, such as corrective hints or explanations, will be more valuable to the extent that users believe the system can target their individual strengths, weaknesses, knowledge, and misconceptions.

Conclusion

W-Pal development and testing have revealed several issues and lessons for building an ITS in ill-defined domains. Some of these feasibility problems may be termed *presentational*, in that they can be overcome by redesigning the interface or mode of instruction to be more modular, engaging, succinct, game-like, and so on. These are relatively easy to fix—more recent iterations of W-Pal have already addressed a number of concerns—although they are often only revealed through extensive usability and feasibility testing. Other feasibility issues may be termed *algorithmic* and relate to the methods by which complex, open-ended, and ambiguous student inputs are processed and evaluated. New and innovative methods for assessing such inputs may be required to realize the full potential of intelligent tutoring in ill-defined domains. However, in ill-defined domains, a certain level of permanent ambiguity may have to be embraced, and the focus must be on guiding students toward progress and independence, rather than delivering, correcting, or testing a well-defined body of knowledge.

References

Aleven, V., & Koedinger, K. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science, 26,* 147–179. doi:10.1207/s15516709cog2602_1

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V. 2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved from http://www.jtla.org

Beal, C., Arroyo, I., Cohen, P., & Woolf, B. (2010). Evaluation of AnimalWatch: In intelligent tutoring system for arithmetic and fractions. *Journal of Interactive Online Learning, 9,* 64–77.

Bell, C., & McNamara, D. (2007). Integrating iSTART into a high school curriculum. *Proceedings of the 29th annual meeting of the Cognitive Science Society* (pp. 809–814). Austin, TX: Cognitive Science Society.

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

Caccamise, D., Franzke, M., Eckhoff, A., Kintsch, E., & Kintsch, W. (2007). Guided practice in technology-based summary writing. In D. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 375–396). Mahwah, NJ: Erlbaum.

Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media, 29,* 241–250. doi:10.1080/1358165042000283101

Crossley, S., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life-Long Learning, 21,* 170–191. doi:10.1504/IJCEELL.2011.040197

Crossley, S., Roscoe, R., Graesser, A., & McNamara, D. (2011). Predicting human scores of essay quality using computational indices of linguistic and textual features. *Proceedings of the 15th international conference on artificial intelligence in education* (pp. 438–440). Auckland, New Zealand: AIED.

Crossley, S., Weston, J., McLain Sullivan, S., & McNamara, D. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28,* 282–311. doi:10.1177/0741088311410188

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

De la Paz, S., & Graham, S. (2002). Explicitly teaching strategies, skills, and knowledge: Writing instruction in middle school classrooms. *Journal of Educational Psychology, 94,* 687–698. doi:10.1037/0022-0663.94.4.687

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*(1), Retrieved from http://www.jtla.org

Flower, L., & Hayes, J. (1981). A cognitive process theory of writing. *College Composition and Communication, 32,* 365–387. doi:10.2307/356600

Gamper, J., & Knapp, J. (2002). A review of intelligent CALL systems. *Computer Assisted Language Learning, 15,* 329–342. doi:10.1076/call.15.4.329.8270

Graesser, A., Lu, S., Jackson, G., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments & Computers, 36,* 180–192. doi:10.3758/BF03195563

Graesser, A., & McNamara, D. (2012). Use of computers to analyze and score essays and open-ended verbal responses. In H. Cooper, P. Camic, R. Gonzalez, D. Long, & A. Panter (Eds.), *APA handbook of research methods in psychology* (pp. 307–325). Washington, DC: American Psychological Association.

Graesser, A., McNamara, D., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point & Query, AutoTutor, and iSTART. *Educational Psychologist, 40,* 225–234. doi:10.1207/s15326985ep4004_4

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology, 104,* 879–896. doi:10.1037/a0029185

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99,* 445–476. doi:10.1037/0022-0663.99.3.445

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.

Hausmann, R., VanLehn, K., Nokes, T., & Gershman, S. (2009). *The design of self-explanation prompts: The fit hypothesis*. Paper presented at the 31st annual meeting of the Cognitive Sciences Society, Amsterdam, the Netherlands.

Hillocks, G. (1984). What works in teaching composition: A meta-analysis of experimental treatment studies. *American Journal of Education, 93,* 133–170. doi:10.1086/443789

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication, 47,* 549–566. doi:10.2307/358601

Jackson, G., & McNamara, D. (2013). Motivation and performance in a game-based intelligent tutoring system. *Journal of Educational Psychology, XX,* XX–XX.

Johnson, W., & Wu, S. (2008). Assessing aptitude for learning with a serious game for foreign language and culture. In B. Woolf, E. Aimeur, R. Nkambo, & S. Lajoie (Eds.), *Intelligent tutoring systems* (pp. 520–529). Berlin, Germany: Springer-Verlag. doi:10.1007/978-3-540-69132-7_55

Kellogg, R., Whiteford, A., & Quinlan, T. (2010). Does automated feedback help students learn to write? *Journal of Educational Computing Research, 42,* 173–196. doi:10.2190/EC.42.2.c

Kintsch, E., Caccamise, D., Franzke, M., Johnson, N., & Dooley, S. (2007). Summary Street®: Computer-guided summary writing. In T. K. Landauer, D. M. McNamara, S. Dennis, & W. Kintsch (Eds.), *Latent semantic analysis* (pp. 263–277). Mahwah, NJ: Erlbaum.

Landauer, T., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice, 10,* 295–308. doi:10.1080/0969594032000148154

Landauer, T., Lochbaum, K., & Dooley, S. (2009). A new formative assessment technology for reading and writing. *Theory Into Practice, 48,* 44–52. doi:10.1080/00405840802577593

Lynch, C., Ashley, K., Pinkwart, N., & Aleven, V. (2009). Concepts, structures, and goals: Redefining ill-definedness. *International Journal of Artificial Intelligence in Education, 19,* 253–266.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke, England: Palgrave. doi:10.1057/9780230511804

McGarrell, H., & Verbeem, J. (2007). Motivating revision of drafts through formative feedback. *ELT Journal, 61,* 228–236. doi:10.1093/elt/ccm030

McNamara, D., Crossley, S., & McCarthy, P. (2010). Linguistic features of writing quality. *Written Communication, 27,* 57–86. doi:10.1177/0741088309351547

McNamara, D., Crossley, S., & Roscoe, R. (2012). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-012-0258-1

McNamara, D. S., O'Reilly, T., Best, R., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research, 34,* 147–171. doi:10.2190/1RU5-HDTJ-A5C8-JVWE

McNamara, D., Raine, R., Roscoe, R., Crossley, S., Dai, J., Cai, Z., . . . Graesser, A. (2011). The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In P. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 298–311) Hershey, PA: IGI Global.

Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability* Retrieved from AQA Centre for Education Research and Policy website: https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf

Michael, J., Rovick, A., Glass, M., Zhou, Y., & Evens, M. (2003). Learning from a computer tutor with natural language capabilities. *Interactive Learning Environments, 11,* 233–262. doi:10.1076/ilee.11.3.233.16543

Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 19,* 139–158. doi:10.1080/10573560308222

Proske, A., Narciss, S., & McNamara, D. (2012). Computer-based scaffolding to facilitate students' development of expertise in academic writing. *Journal of Research in Reading, 35,* 136–152. doi:10.1111/j.1467-9817.2010.01450.x

Rock, J. (2007). *The impact of short-term use of Criterion on writing skills in 9th grade* (Research Report no. RR-07–07). Princeton, NJ: Educational Testing Service.

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

Roscoe, R., Varner, L., Weston, J., Crossley, S., & McNamara, D. (in press). The Writing Pal Intelligent Tutoring System: Usability testing and development. *Computers and Composition*.

Rowley, K., & Meyer, N. (2003). The effect of a computer tutor for writers on student writing achievement. *Journal of Educational Computing Research, 29,* 169–187. doi:10.2190/3WVD-BKEY-PK0D-TTR7

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

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). Cambridge, MA: AAAI Press/MIT Press.

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

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

Shute, V. (2008). Focus on formative feedback. *Review of Educational Research, 78,* 153–189. doi:10.3102/0034654307313795

Simon, H. (1973). The structure of ill structured problems. *Artificial Intelligence, 4,* 181–201. doi:10.1016/0004-3702(73)90011-8

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J., Shelby, R., Taylor, L., . . . Wintersgill, M. (2005). The Andes Physics Tutoring System: Lessons learned. *International Journal of Artificial Intelligence in Education, 15,* 147–204.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10,* 157–180. doi:10.1191/1362168806lr190oa

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

Wolfe, C., Britt, M., Petrovic, M., Albrecht, M., & Kopp, K. (2009). The efficacy of a web-based counterargument tutor. *Behavior Research Methods, 41,* 691–698. doi:10.3758/BRM.41.3.691

Received December 15, 2011
Revision received December 18, 2012
Accepted February 11, 2013 ■