Available online at www.sciencedirect.com

ScienceDirect

Computers and Composition 34 (2014) 39–59

Computers and Composition

www.elsevier.com/locate/compcom

The Writing Pal Intelligent Tutoring System: Usability Testing and Development

Rod D. Roscoe [a,*], Laura K. Allen [b], Jennifer L. Weston [b], Scott A. Crossley [c], Danielle S. McNamara [d]

[a] *Human Systems Engineering in the Ira A. Fulton Schools of Engineering at Arizona State University, United States*
[b] *Psychology and the Learning Sciences Institute at Arizona State University, United States*
[c] *Applied Linguistics and ESL at Georgia State University, United States*
[d] *Learning Sciences Institute and Professor of Psychology at Arizona State University, United States*

Abstract

The Writing Pal (W-Pal) is an intelligent tutoring system (ITS) designed to improve students' writing proficiency via a unique combination of explicit strategy instruction, game-based practice, essay writing practice, and automated formative feedback. To develop and refine the many features of the W-Pal tutoring system, we have employed a multiple-method usability testing approach, which capitalizes on the complementary strengths and weakness of methods such as focus groups, component studies, internal testing, and in vivo testing. These diverse methods allow researchers to benefit from focused student input, instructor input, and iterative development, while also gathering data in ecologically-valid settings. In this paper, we describe some of the testing and development of aspects of W-Pal, consider the challenges of building such a system, and provide a particular emphasis on a feasibility study that integrated W-Pal into high school English classrooms during a school year. The results of this study showed that students perceived the system as informative, valuable, and enjoyable, and results also highlighted specific ways that these aspects of the system could be further enhanced. Based upon these findings, a significantly updated version of W-Pal has been developed. The current W-Pal system is described along with considerations for future research and how the system may be used to supplement writing instruction.

Keywords: Writing instruction; Writing strategies; Intelligent tutoring systems; Automated writing evaluation; Formative feedback; Educational games; Usability testing

1. The Writing Pal intelligent tutoring system: Usability testing and development

The Writing Pal (W-Pal) is an intelligent tutoring system (ITS) designed to improve students' writing proficiency via a unique combination of strategy instruction, game-based practice, essay writing practice, and automated formative feedback. This novel technology has been developed over several years of iterative, interdisciplinary work that synthesizes best practices from writing pedagogy and user-centered design (Dai, Raine, Roscoe, Cai, & McNamara, 2011; McNamara et al., 2012), with contributions from psychology, composition, linguistics, and computer science.

* Corresponding author.
E-mail address: rod.roscoe@asu.edu (R.D. Roscoe).

http://dx.doi.org/10.1016/j.compcom.2014.09.002
8755-4615/

This work incorporates design principles common to many ITSs, such as cross-platform accessibility (i.e., PC and Mac) at school and at home, and database structures that enable robust data logging and continuation of service despite interruptions.

W-Pal has also been informed by fundamental pedagogical design principles. First, W-Pal was designed with the assumption that writing is a complex process with several non-linear phases, such as prewriting, drafting, and revising (e.g., Deane et al., 2008; Flower & Hayes, 1981). Numerous strategies can be taught to support these phases (e.g., Graham & Perin, 2007; Hillocks, 1984), and teachers need freedom to utilize W-Pal in ways that meet diverse curricula. Thus, W-Pal presents strategies via multiple *Strategy Modules* that can be studied and practiced in any order. Second, the development of writing proficiency requires opportunities for sustained practice, yet student engagement with ITSs can decrease over time, leading to disengagement (Bell & McNamara, 2007). To ameliorate these effects, W-Pal offers numerous educational games that enable strategy practice, which are hypothesized to foster motivation (Barab, Thomas, Dodge, Carteaux, & Tuzun, 2005; McNamara, Jackson, & Graesser, 2010). Third, writing development requires individualized and formative feedback (McGarrell & Verbeem, 2007; Sommers, 1982). This is a significant challenge because writing and assessment are highly subjective (Huot, 1996; Meadows & Billington, 2005). W-Pal, like human instructors, must assess open-ended input, including essays and writing excerpts, and provide meaningful responses. For this purpose, we have compiled a number of natural language processing (NLP) tools to inform our writing assessment algorithms. These tools include a lemmatizer, syntactic parsers, lexical databases (Princeton University's WordNet, the Max Planck Institute for Psycholinguistics's CELEX, and the MRC Psycholinguistic Database), rhetorical analyzers, and Latent Semantic Analysis (LSA; Graesser & McNamara, 2011; McNamara & Graesser, 2012). NLP algorithms also drive feedback in W-Pal. This, too, has been challenging because there is little empirical research on how to design and implement formative feedback based on automated assessments (Grimes & Warschauer, 2010; Roscoe, Kugler, Crossley, Weston, & McNamara, 2012; Roscoe et al., 2011; Roscoe, Varner, Crossley, & McNamara, 2013).

W-Pal is not alone in providing automated writing assessment and feedback. A number of automated writing evaluation systems (AWEs) are available. These programs score essays and offer feedback via a combination of statistical modeling, NLP, LSA, machine learning, and other artificial intelligence methods. Systems such as Education Testing Service's Criterion (Attali & Burstein, 2006; Burstein, Chodorow, & Leacock, 2004) and Vantage Learning's MY Access! (Grimes & Warschauer, 2010; Rudner, Garcia, & Welch, 2006) rely primarily on statistical modeling. In such systems, an essay corpus is annotated to identify essay elements (e.g., topic sentences). Essays are then analyzed along many linguistic dimensions, and statistical analyses extract features that discriminate between higher- and lower-quality essays. Statistical models combine extracted linguistic properties into algorithms that assign grades to essays. In contrast, the Pearson Intelligent Essay Assessor (Landauer, Laham, & Foltz, 2003; Wohlpart, Lindsey, & Rademacher, 2008), uses LSA to assess essays. LSA assumes that similar words occur in similar contexts (e.g., sentences, paragraphs, or whole texts). Singular value decomposition is used to reduce a large word-by-context matrix to the number of dimensions needed to capture semantic structure. Essay scores are based on the semantic similarity between student essays and a benchmark corpus of pre-scored essays.

Proponents of AWEs argue students receive more opportunities to write and receive feedback without adding to instructors' workload. Providing timely, personalized feedback on student writing is rewarding but time-intensive, and AWEs allow teachers to assign more writing than might otherwise be feasible. However, two objections to AWEs are that they lack human sensitivity and that the classification of writing quality is limited by available algorithms (Hearst, 2002). AWEs, with their reliance on broad statistical regularities, may not capture writers' unique individual expressive differences (Clauser, Kane, & Swanson, 2002). In addition, savvy users can potentially trick the scoring system (Powers, Burstein, Chodorow, Fowles, & Kukich, 2002). Thus, despite progress, automated scoring systems are still under development with areas for improvement. Further information about the history and development of such scoring tools can be found in overviews by Mark Warschauer and Paige Ware (2006) and Semire Dikli (2006).

The fundamental differences between W-Pal and AWEs reside in their respective origins and pedagogy. AWEs such as MY Access! began as scoring systems (e.g., the core of MY Access! is the Vantage Learning IntelliMetric scoring engine). The purpose of such tools was to facilitate the rating of essay quality. Indeed, much of the AWE literature focuses on demonstrating scoring reliability and accuracy (Grimes & Warschauer, 2010). Over time, algorithms increased in sophistication and could address finer-grained essay traits that could then be communicated to students as feedback. In short, the conceptualization of AWEs began with scoring and assessment but has seen a gradual accumulation of pedagogical utility.