Available online at www.sciencedirect.com

ScienceDirect

Computers and Composition 42 (2016) 1–16

Computers and Composition

www.elsevier.com/locate/compcom

A critical interpretative synthesis: The integration of Automated Writing Evaluation into classroom writing instruction

Marie Stevenson

University of Sydney, Australia

Abstract

Automated Writing Evaluation (AWE) is computer-generated scoring and feedback that is used for both assessment and instructional purposes. Much controversy has surrounded AWE, especially in high-stakes tests such as TOEFL, and much of the discussion has centered around the scoring and feedback capabilities of AWE and the effects of AWE on text quality. Relatively little attention has been directed towards the ways that AWE is used or could be used as an instructional tool in the writing classroom. Through a critical interpretative synthesis of existing research, this study provides an overview of what is currently known about the integration of AWE into classroom writing instruction. The synthesis found that that there are numerous purposes for using AWE stated in existing research, some of which do not accord with objectives commonly associated with AWE; that teachers had varied and creative ways of integrating AWE in their classrooms; and that, although students generally seemed to enjoy using AWE, at the times when the sample studies were conducted, there appeared to be many limitations in the feedback provided by AWE systems. The study discusses these findings in terms of criticisms that have been leveled against AWE and links this discussion to broader considerations of the relationship between literacy, technology and pedagogy.

Keywords: automated writing evaluation; AWE; feedback; writing instruction; literacy

1. Introduction

Automated Writing Evaluation (henceforth AWE) involves computer-generated scoring and feedback for writing. Numerous commercial and non-commercial AWE systems are available, the central component of which is a scoring engine that generates automated scores based on techniques such as artificial intelligence, natural language processing and latent semantic analysis. Many AWE systems also incorporate written feedback on various aspects of writing. However, most AWE systems model only a relatively small part of the writing construct, being largely concerned with structure (e.g., topic sentences and paragraph transitions); phrasing (e.g., vocabulary and sentence length); and transcribing (e.g., spelling and mechanics (Deane, 2013).

AWE was originally developed to generate summative scores for assessment purposes, and is currently being used, in combination in human evaluation, in high-stakes tests such as the Test of English as a Foreign Language (TOEFL) and the Graduate Management Admissions Test (GMAT). However, the use of AWE feedback as an instructional tool in writing classrooms is increasing, especially in school and college classrooms in the United States.

E-mail address: marie.stevenson@sydney.edu.au

http://dx.doi.org/10.1016/j.compcom.2016.05.001
8755-4615/

Considerable controversy has surrounded AWE, particularly its use in high stakes testing situations. This controversy has centered around doubts concerning the accuracy of scoring and feedback capabilities, and fears concerning the effects of writing for a non-human audience. There have also been criticisms concerning the restricted and objectivistic views of writing and assessment that are said by some to underpin AWE scoring and feedback. Vojak, Kline, Cope, McCarthey, & Kalantzis (2011) examined scoring and feedback features in terms whether AWE systems harness the potential of new technologies to promote new literacies. The authors found that AWE systems generally failed to reflect social, contextual and multi-modal aspects of writing.

AWE validation research, much of which has been carried out by researchers affiliated in some way with companies that develop and market AWE systems, has focused on the psychometric properties of AWE scoring, with the objective of demonstrating that AWE systems score as reliability and validly as human raters. AWE pedagogical research has predominantly focused on establishing whether AWE scoring and feedback have positive effects on the quality of texts produced by student writers. A critical review of research on the effects of AWE on text quality (Stevenson & Phakiti, 2014) found that there was relatively little attention in the literature as to how AWE was used in the classroom, or to how it could be effectively integrated into classroom instruction.

Nonetheless, a small number of studies do examine use of AWE in the classroom, and some of the research that focuses on the effects of AWE on text production does include a modest amount of discussion of classroom use. Through a critical interpretative synthesis of existing research, the current study provides an overview of what is currently known about the integration of AWE into classroom writing instruction. It examines three commercially available web-based AWE systems: Criterion, MY Access! and Summary Street.

The study does not aim to describe or compare specific AWE programs. Rather, by examining classroom integration, it seeks to go beyond providing descriptive "gee-whiz explanations of new technologies" (Luke & Luke, 2001: 93). The study discusses the findings in terms of criticisms that have been leveled against AWE. This discussion is linked to broader considerations of the relationship between literacy, technology and pedagogy.

2. The controversy surrounding AWE

The controversy surrounding AWE has been widespread. An online petition, "Professionals Against Machine Scoring of Student Essays in High-Stakes Assessment" received thousands of signatures, including Noam Chomsky's, and was cited in a number of newspapers, including *The New York Times.* The machine-scoring of writing for assessment purposes has also been opposed by the Conference on College Composition and Communication (CCCC) in *Position statement on teaching, learning, and assessing writing in digital environments* (2004) and *Writing assessment: A position statement* (2009). The 2004 statement states that: "Writing-to-a-machine violates the essentially social nature of writing: we write to others for social purposes." In 2006, Patricia Freitag Ericsson & Richard Haswell edited a book entitled *Machine Scoring of Student Essays: Truth and Consequences*, which consisted of a series of papers strongly questioning the purported accuracy of computerized essay scoring, warning of the dehumanization of writing caused by writing for a machine, and decrying the use of AWE to replace human raters in testing situations.

It is tempting to dismiss some of these criticisms as a kind of neo-Luddism, and indeed this charge has been made by those who point out that the teaching practices these criticisms defend are the same ones that were vehemently attacked with their introduction in the industrial revolution (e.g., Elliot & Klobucar, 2013). Certainly, there is a technophobic tone in some of the criticisms, for example, the introduction of the Freitag Ericsson & Haswell book:

> "...new technology can sneak in the back door and establish itself while those at the front gates, nominally in charge, are not much noticing. All of a sudden cell phones are disturbing legislative sessions and church services and allowing students to cheat on examinations in new ways. All of a sudden students can pass entrance examination essays in ways never allowed before, with their essays scored by machines running commercial software programs" (p. 1).

However, criticisms of AWE have also been voiced by those who champion technology as a dominant shaping force in what are referred to as 'new literacies'. (See Buckingham (1993) and Lankshear and Knobel (2006) for discussion of new literacies). Much has been made of properties of new literacies such as multi-modality, synchronicity (i.e., real-time on-line communication) and non-linearity (e.g., hypertexts), and the need to incorporate new literacies into