Book review

Shermis, M.D. & Burstein, J. (Eds) (2013). Handbook of Automated Essay Evaluation: Current applications and new directions. Routledge: New York and London.
ISBN**-10: 0415810965**

Reviewed by: Marie Stevenson, Universtiy of Sydney

This book is about Automated Essay Evaluation (AEE), which is computer-generated scoring and written feedback on the quality of written texts. Numerous commercial and non-commercial AEE programs are currently 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. AEE scoring engines are trained using human ratings, and AEE is being used for summative assessment in tests such as the Test of English as a Foreign Language (TOEFL) and the Graduate Management Admissions Test, in combination with human ratings. Many AEE programs now also provide written feedback in the form of comments and corrections. In recent years, the use of written AEE feedback for the provision of formative feedback in writing classrooms in both schools and colleges has increased, particularly in classrooms in the United States.

The book follows on from a previous volume published in 2003, titled: *Automated Essay Scoring: A Cross-Disciplinary Approach* (Lawrence Erlbaum Associates, Inc., 2003). It aims to provide a comprehensive overview of developments in the field that have occurred in the last ten years. According to the foreword, the shift in nomenclature between the volumes from Automated Essay Scoring to Automated Essay Evaluation is a significant one. Reflecting the assessment-oriented origins of AEE alias AES, the focus of the previous volume was on scores generated for testing purposes, and on the software that generated these scores. It had a strong computational-linguistic and psychometric focus. The foreword appears to set the stage for the latest volume to broaden its focus by including written AEE feedback designed for teaching purposes and also by including voices from educational policy and writing research, in addition to those of computational linguists and psychometricians.

journal of
WRITING RESEARCH

Shermis, M.D. & Burstein, J. (Eds) (2013). Handbook of Automated Essay Evaluation: Current applications and new directions. [Book Review by Marie Stevenson]. *Journal of Writing Research, 5*(2), 239-243. http://dx.doi.org/10.17239/jowr-2013.05.02.4

Contact: Marie Stevenson, University of Sydney, Faculty of Education and Social Work, Camperdown Campus, Sydney, 2006, Australia– marie.stevenson@sydney.edu.au

The book positions AEE as controversial. In fact, the author of the foreword goes as far describing himself as being "terrified" about writing the foreword, due to "the drumbeat of criticism about AEE"(p vii). Perhaps not such an exaggeration, given that earlier this year, an online petition, "Professionals Against Machine Scoring of Student Essays in High-Stakes Assessment" was signed by thousands of people, including Noam Chomsky, and was cited in a number of newspapers, including *The New York Times*. The foreword acknowledges awareness of the barrage of criticism of AEE by writing teachers and researchers about issues such as its ability to provide accurate, cheat-proof scores; the possible effects of writing for a non-human audience; and the formalistic nature of the feedback provided. Thus, the stage also appears to be set for this volume to come to grips with these issues by providing an examination of the existing evidence for the effectiveness of AEE in relation to students' writing, and by providing an insightful and well-balanced discussion.

Synopsis

The preface informs us that the book is loosely divided into four sections. The first chapter provides an overall introduction to AEE by describing AEE and providing a history of its development, and outlining some of the concerns voiced about AEE. Below follows a brief synopsis of each of the four sections. As neither the preface nor the table of contents informs the reader where the sections begin and end, I have used my own discretion here.

Section 1

Section 1 is about AEE and writing research. Chapter 2 is about the use of AEE as a rapid assessment tool to diagnose first year college students in the US in need of remedial writing instruction. The chapter advocates the use of AEE for this purpose, citing both resulting decreases in remediation rates and significant positive correlations between AEE scores, SAT scores and writing portfolio scores as justifications. Chapter 3 provides an evaluation of the pros and cons of AEE scoring and feedback. Although its focus is on EFL and ESL writing, many of its insights concerning the capabilities or lack thereof of AEE are equally applicable to writing in general. The chapter also contains a brief section on whether automated feedback is effective in improving students' writing.

Section 2

Section two is about the characteristics and capabilities of specific AEE programs. The chapters in this section have a strongly computational linguistic and psychometric orientation, and the main focus is on scores rather than written feedback. Chapter 4 is about E-rater, developed by Education Testing Service, which uses NLP to identify linguistic properties in a text. Chapter 5 is about Intelligent Essay Assessor, developed by Pearson Knowledge Technologies, which is based on Latent Semantic Analysis. Chapter 6 is about Intellimetric, developed by Vantage Learning, which uses hybrid

techniques, including linear analysis, Bayesian and Latent Semantic analysis; Chapter 7 is about a summative (WESTEST 2) and a formative assessment program (West Virginia Write) being implemented in schools in West Virginia. Chapter 8 is about LightSIDE, originally built at Carnegie Mellon's Language Technologies Institute, which uses what is known as open source technology.

Section 3

Section 3 deals largely with psychometric issues surrounding AEE, and again deals with scoring rather than written feedback. Chapter 9 examines the challenges and logistics of automated short answer scoring. Chapter 10 examines the validity of AEE by considering how the reasoning that supports the assignment of scores to essays by human raters might be different when automated scoring is used. Chapter 11 outlines a framework for establishing the validity and reliability of AEE scoring, and concludes that the evidence so far indicates that AEE is able to measure some, but not all, aspects of writing. Chapter 12 is about scaling and norming for automated essay scoring, and is a highly psychometric, statistically-oriented chapter. Chapter 13 considers the relationship between AEE scores and human ratings. Chapter 14 considers the phenomenon of 'reader drift', which concerns the tendency of human raters to drift away from scoring criteria, and considers the ways in which AEE can be utilized to monitor human rater performance.

Section 4

Section 4 is about a variety of current developments in the field of AEE, namely, using AEE to evaluate discourse coherence in essays (Chapter 15); current techniques in AEE grammar error detection (Chapter 15); using AEE to evaluate discourse coherence (Chapter 16); using AEE to identify attitudinal expressions (Chapter 17); and using a cognitive model to identify skills and processes relevant to the writing construct in order to develop AEE systems to incorporate these measures (Chapter 18). Chapter 19 compares AEE systems in terms of the relationships between AEE scores and human scores. The book concludes with a chapter that is more educationally-oriented, examining the role that AEE can play in the Common Core State Standards Initiative (CCSSI), which is a common set of education K-12 standards in Language Arts/Literacy and Mathematics in the US.

Evaluation

Overall, this book provides an impressive, if sometimes weighty, compendium of the current state of development of AEE systems. For those interested in deepening their understanding of how AEE systems work, the capabilities these systems have, and how these capabilities compare to those of human raters or to those of other AEE systems, the book is of immense value. The book provides an excellent overview of the major developments that have taken place in the last ten years, right up to the cutting edge. It

also provides in-depth discussion of key issues such as validity, reliability, and norming that lie at the heart of developing accurate and meaningful automated scoring systems.

However, unfortunately, the book does not really fulfill its promise of providing a broader, more inclusive perspective on AEE that includes voices from writing research and educational policy. Like the previous volume, the book seems to be largely written for computational linguistic and psychometric audiences. This is likely because most of the contributing authors come from these backgrounds and, as the list of contributors indicates, many are employed by or are closely affiliated the companies that develop and market AEE systems.

The writing researcher or educationalist negotiating the book may find him or herself puzzled and perhaps a little disappointed that so few links are made to theoretical and pedagogical principles derived from writing research or writing instruction. As the focus remains on summative rather than formative evaluation, the reader is given little or no insights into how AEE written feedback is being used formatively in classrooms, what its effects on writers are, or how it can be integrated effectively into classroom instruction. Of the two writing instruction chapters, one (i.e. Chapter 2) turns out to be about the use of AWE for assessment purposes, and has little or nothing to say about the use of AWE in the writing classroom. The other (i.e. Chapter 3), which is by a long shot most informative from a writing research/instruction perspective, does incorporate both summative and formative uses of AEE. It should also be said that Chapter 18, though not pedagogically-oriented, does link the further development of AEE to a cognitive model of writing, and through this to specific skills and processes, which is a very welcome development.

The focus of the book is on AEE systems themselves, rather than on writing or writers. However, many of the negative claims about AEE have centred around its purported effects on writers and writing. The book is honest and open in acknowledging the limitations that AEE systems are grappling with in measuring aspects of the writing construct, but by largely leaving writers out of the equation it is by definition unable to weigh up, let alone counter, many of the criticisms that have been made. It provides only a very superficial exploration of the effects of AEE on students' writing. Two of the chapters (Chapter 3 & Chapter 15) touch on the effects of AEE feedback on students' writing, but both stop short of discussing these in any depth, or giving anything that really resembles a review of the literature on this topic. This is to be lamented, as a growing body of literature exists that considers the effects of AWE terms of the quality of students writing, the effects of AEE on students' writing processes, and the effects on student and teacher perceptions, and classroom use of AEE systems. Most of this literature is not referred to either in this chapter or elsewhere in the book.

All in all, in terms of being a well-rounded handbook of automated essay evaluation, the book promises more than it is able to deliver. However, if the reader ignores the half-kept promises and accepts the book for what it largely is – a sound,

rigorous and in-depth treatment of AEE scoring systems, their capabilities, their psychometric properties, and their ongoing development – then the reader will be well rewarded by this book.