writing & pedagogy

ISSN: 1756–5839 (print)
ISSN: 1756–5847 (online)

New Books

Handbook of Automated Essay Evaluation
Current Applications and New Directions
Mark D. Shermis and Jill Burstein (eds.) (2013)

New York: Routledge. Pp. 194
ISBN: 9780415810968

Reviewed by Jane Lockwood

This edited collection of highly readable chapters on different aspects of the automated assessment of writing is a timely collection, particularly for the uninitiated. For too long issues related to automated language assessment have been the concern of those steeped in the fields of language testing and computational linguistics. This situation is however now changing. Educational institutions, governments, and businesses worldwide are increasingly requiring more and faster information about the language communication skills of their students, citizens, and employees; and this means many more of us need to become comfortable with testing and assessment theory and how technology can assist our efforts in evaluating writing.

Automated language assessment is fast becoming a part of our digital teaching and learning world, and its possibilities have grown exponentially over the last decade. This can be threatening for some language teachers, as computational linguistics, mathematical models, and the building of algorithms for technology-driven solutions are usually beyond their knowledge and skill set. Teachers recognize that writing is perhaps the most resource-intensive English language skill to assess, and they could do with some help; but they recoil at the idea of a machine being the main audience for their students' writing or fear that language assessment technology will

Affiliation
City University of Hong Kong
email: lockwood@cityu.edu.hk

WAP VOL 6.2 2014 437–441

doi: 10.1558/wap.v6i2.437

marginalize and de-professionalize them. All in all, automated language assessment has had a bad press.

However, with the demands of governments, of globalized English-speaking workplaces, and of English-medium of instruction (EMI) universities and schools worldwide, we are compelled to seek improvements in the assessment of writing, including automated essay evaluation (AEE).The new "policy landscape" of the Common Core Standards Initiative (CCSI) in the United States is comprehensively reviewed in Chapter 20 ("The Policy Turn in Current Education Reform," by Kenji Hakuta) with specific reference to the implications (e.g. the grading of millions of school essays) for educational practice using automated assessment. It would seem that technology, after all, may have something to offer, and the set of chapters in the Shermis and Burnstein volume certainly supports this position. Written by experts in automated writing assessment systems, the collection opens up the world of AEE with clarity and practicality. It enables us to begin to understand how automated assessment might assist, rather than replace teachers in writing classes.

An important distinction is made at the beginning of the book between *automated essay scoring* (AES) and *automated essay evaluation* (AEE). This distinction mirrors an important recent shift in the focus of language assessment from summative scoring for reporting purposes (assessment of learning) to formative feedback for instructional purposes (assessment for learning). Many of the chapters are sensitive to this shift and explain how many automated systems are now capable of providing valid and reliable diagnostic feedback. In a practical sense, of course, the attraction of automated systems is obvious. They cut down human rater time dramatically; they never get tired, and therefore grading turnaround time is significantly reduced; and machines, while no doubt expensive to set up, do not need to be paid on an hourly basis or by script. But do we ultimately trust them?

Mark Shermis, Jill Burnstein, and Sharon Apel Bursky (Chapter 1, "Introduction to Automated Essay Evaluation") offer a lively and highly informative opening to the book by addressing some of the concerns of teachers outlined above and presenting an overview of the development of technology to support AEE. They propose a simple definition of AEE as "an ability of computer technology to evaluate and score written prose" (Shermis and Burnstein, 2003: xiii) and explain that computers work to algorithms that have been programmed to replicate the scores given to the same essays by experienced and reliable human markers. They cite two key computer-based approaches commonly used; the first is Natural Language Processing (NLP), which uses techniques to identify specific lexical and syntactic cues for text analysis and is specifically exemplified

in Chapter 4 ("The E-Rater Automated Essay Scoring System," by Jill Burstein, Joel Tetreault, and Nitin Manini) by the "e-rater" system. The other approach is Latent Semantic Analysis (LSA), which focuses on content-related features and can include diagnostic feedback on grammar, style, and mechanics of writing, as explained in Chapter 5 ("Implementation and Applications of the Intelligent Essay Assessor," by Peter Foltz, Lynn Streeter, Karen Lochbaum, and Thomas Landauer) and Chapter 6 ("The Intellimetric Automated Essay Scoring Engine – A Review and an Application to Chinese Essay Scoring," by Matthew Schultz). The different linguistic analyses described above suggest salient features or criteria that human raters may employ in using a rubric; these features, or "proxies," are then identified in the humanly marked scripts and converted into algorithms. In other words, the algorithms are only as good as the rubric criteria and the reliability of the human assessors. AEE does not, it would seem, try to outdo teachers as much as try to mimic them.

Of particular note is Shermis, Burnstein, and Bursky's view that the development of AEE technology is dependent on writing teachers, test developers, cognitive psychologists, psychometricians, and computer scientists working collaboratively. Furthermore, they note with concern that most AEE technology systems are not available publicly, which to some extent limits experimentation and collaboration; but they single out the LightSIDE application (Chapter 8, "LightSIDE: Open Source Machine Learning for Text," by Elijah Mayfield and Carolyn Penstein Rose), which they describe as an "easy to use automated evaluation engine with both compiled and source code publicly available" (p. 11).

Of specific relevance to English as a second language (ESL) teachers is Sara Weigle's chapter (Chapter 3, "English as a Second Language Writing and Automated Essay Evaluation"), which is focused on the pedagogical concerns of ESL teachers and considers how these might be dealt with using automated assessment. She addresses the issues of the theoretical constructs and pedagogical approaches underpinning such practices and is candid in her comments in cases where she feels there is, and is not, a contribution to be made. For example, she suggests that there may be more value in using automated assessment for lower proficiency than higher proficiency ESL students, when, for example, the latter group may be working more on aspects of rhetorical effectiveness and audience persuasion. She also touches on the issue of using automated assessment in high-stakes versus low-stakes assessment contexts. In a similar vein, Norbert Elliot and Andrew Klobucar (Chapter 2, "Automated Essay Evaluation and the Teaching of Writing) mediate the advantages of new technologies in automated assessment while safeguarding the integrity

of current writing pedagogy and practice; and Chris Brew and Claudia Leacock explore the possibility in Chapter 9 ("Automated Short Answer Scoring: Principles and Prospects) of using automated assessment to rate short answer questions rather than essays with some early positive results.

The important, yet difficult, issue of establishing validity of AES and AEE is discussed in some detail in both Chapter 10 ("Probable Cause: developing Warrants for Automated Scoring of Essays," by David Williamson) and Chapter 19 ("Contrasting State-of-the-Art Scoring of Essays," by Mark Shermis and Ben Hamner), where the authors of each chapter define validity in different ways in order to interrogate this issue. A further angle on validity is provided in Chapter 14 ("Using Automated Scoring to Monitor Reader Performance and Detect Reader Drift in Essay Scoring"), where Susan Lottridge, E. Matthew Schulz, and Howard Mitzel explore how both human and automated scoring can convincingly be used together in a "blended" solution. Finally, Kristin Koskey and Mark Shermis (Chapter 12, "Scaling and Norming for Automated Essay Scoring") tackle the process of scaling and norming essay scores to produce meaningful AES scores by taking us back to general principles. Another contribution of note discusses the application AES and AEE in real settings, in Chapter 7 ("Applications of Automated Essay Evaluation in West Virginia," by Changua Rich, M. Christina Schneider, and Juan M. D'Brot), where experimentation in using both summative and formative automated assessment tools is discussed.

In a very real sense, the collection is not only a good introduction to automated systems for writing assessment but also a good introduction to those who may be new to issues and principles in writing and to assessment constructs and principles of assessment validity, reliability, and practicality. Several of the articles (see Chapters 11, "Validity and Reliability of Automated Essay Scoring," by Yigal Attali; Chapter 13, "Human Ratings and Automated Essay Evaluation," by Brent Bridgeman; and Chapter15, "Grammatical Error Detection in Automatic Essay Scoring and Feedback," by Michael Gamon, Martin Chadorow, Claudia Leacock, and Joel Tetreault) start off by interrogating the theoretical constructs for writing in general, and then proceed to highlight what they see as important differences between human and automated rating. In Chapter 16 ("Automated Evaluation of Discourse Coherence Quality in Essay Writing"), for example, Jill Burstein, Joel Tetreault, Martin Chodorow, Daniel Blanchard, and Slava Andreyev provide an introduction to the linguistic and psychological constructs for writing before introducing their particular approach for the development of a new automated tool. Pedagogy underpins the articulation of the constructs upon which assessment frameworks are derived, and there is a vein of argumentation throughout the book that contests traditionally

derived criteria for writing. In Chapter 18, Paul Deane, in "Covering the Construct," engages with the issue of applying the traditional constructs used by human raters to assess literacy and contemplates a shift to "using a general cognitive framework as a construct model for writing" (p. 298), where, in the future, we will rely solely on technology. This discussion culminates in the contribution by Jill Burstein, Beata Beigman-Klebanov, Nitin Manini, and Adam Faulkner (Chapter 17, "Automated Sentiment Analysis for Essay Evaluation"), where "sentiment analysis" identifies polarity stances (positivity, negativity, and neutrality) in student writing.

Technologically, two themes emerge from this excellent collection. First is that the technology and mathematical modeling and derivation of algorithms reflecting the construct (or rather bits of the construct) of writing is a highly complex process. What many of the chapters do is attempt to make this accessible to all of us non-testing specialists. Are they successful in doing this? For me, like many other readers with limited statistical and mathematical backgrounds, there was enough detail to see that the algorithm is only ever going to be as good as the construct and the human rater output behind it. The second theme relates to the future inevitability of AES and AEE use and the need to work with it rather than against it. In my view, the contributions open up the possibilities of automated writing assessment and embrace the complexity, rather than closing it down. I would highly recommend this book to all first and second language writing teachers; the possibilities for real support in automated writing assessment are persuasive.

About the Author

Jane Lockwood is Associate Professor in the Department of English at City University of Hong Kong and is currently leading a language assessment research project focusing on automated writing assessment for undergraduate students studying in Hong Kong universities. She holds a Ph.D. in Applied Linguistics from Hong Kong University and has published widely in English for Specific Purpose (ESP) curriculum development and performance language assessment at workplaces and universities. She is a Guest Editor for a Special Topic Issue of *Writing and Pedagogy* on Writing Assessment, 7(3), to be published in Winter 2015.

Reference

Shermis, M. D. and Burstein, J. (2003) Introduction. In M. D. Shermis and J. Burstein (eds.), *Automated Essay Scoring: A Cross Disciplinary Perspective* xiii-xvi. Mahwah, New Jersey: Lawrence Erlbaum Associates.