June 2014, Volume 18, Number 2
pp. 65–69

REVIEW OF *HANDBOOK OF AUTOMATED ESSAY EVALUATION: CURRENT APPLICATIONS AND NEW DIRECTIONS*

Handbook of Automated Essay Evaluation: Current Applications and New Directions

Mark D. Shermis & Jill Burstein (Eds.)

2013

ISBN: 978-1-84872-995-7 (hard cover)

ISBN: 978-0-415-81096-8 (paperback)

ISBN: 978-0-203-12276-1 (e-book)

US $295 (hard cover)

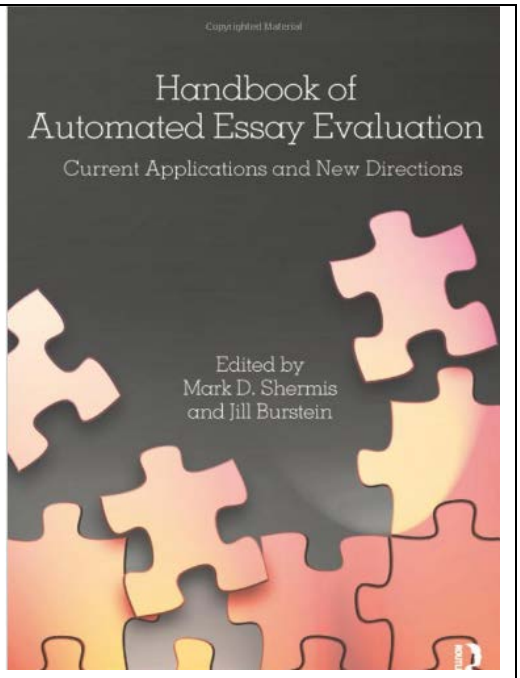
US $115 (paperback)

US $109.85 (e-book)

362 pp.

Routledge

New York



Review by Li Zhang, Shanghai Jiao Tong University

Since Ellis Page (1966) published his landmark article, "The Imminence of Grading Essays by Computer," the issue of using computers to grade writing and provide feedback has been a great concern for researchers in language testing and writing instruction. Research has been done for the creation and development of systems for automated essay evaluation (AEE), or automated essay scoring (AES), what has been called the process of evaluation and scoring written prose via computer programs (Shermis & Burstein, 2003). The *Handbook of automated essay evaluation* edited by Mark D. Shermis and Jill Burstein collects the most recent articles about research on AEE and provides comprehensive perspectives for addressing the following issues: AEE and writing instruction, applications of different AEE engines, validity and reliability of AEE, and discussions of aspects reflecting writing constructs in AEE. This handbook, therefore, is beneficial for both researchers and writing instructors who are interested in investigating how AEE can be better applied in essay assessment and writing instruction.

In the introductory chapter, Shermis, Burstein and Bursky firstly state the major concerns about AEE, namely that computers cannot evaluate essays as well as humans can. Thus they suggest a multidisciplinary approach that involves cognitive psychology and psychometric evaluations. After reviewing the early history of the application of AEE, they generalize the technology that facilities the use of AEE, such as word processing, the Internet and natural language processing (NLP), and briefly describe the evolution of commercial AEE. Finally, the authors present guidelines for how the following chapters are organized.

The next two chapters deal with AEE and the teaching of writing. Following a literature review of research on teaching and assessing writing, Elliot and Klobucar, in Chapter 2, examine models of writing constructs and report case studies conducted over a three-year period at a science and technology university. Based on previous research as well as on their own work, they conclude the chapter by

pointing to directions for future developments in writing assessment.

Chapter 3 treats AEE and writing instruction for non-native speakers. Weigle starts with a review of the context for assessing non-native writing and following the examination of the construct of writing intended for learners of different writing proficiencies, she briefly discusses the two major functions of AEE—scoring and feedback—and gives suggestions on how to implement AEE effectively in different contexts, such as giving teachers training and support, helping students interpret feedback, and presenting the tools carefully. She states firmly in the conclusion that failures in the implementation of educational technology do not result from problems of technology but rather from resistance from teachers and thus urges teachers to be part of the "steamroller" (p. 51) for the use of AEE.

Chapters 4 through 9 offer a number of articles addressing features and functions of various AEE systems. Burstein, Tetreault and Madnani, in Chapter 4, examine how the application of e-rater fits into problem spaces in curriculum and assessment development. They begin with an introduction of features of e-rater and their relevance to writing construct. Then, they explain how NLP methods are employed to identify construct-relevant linguistic properties in text, which include the statistical and rule-based methods. Next, they describe the kinds of e-rater features that use an NLP approach. They finally highlight the importance of relating e-rater development to language requirements specified in the Common Core State Standard Initiative in the United States.

Intelligent Essay Assessor (IEA) is introduced by Foltz, Streeter, Lochbaum and Landauer in Chapter 5. They describe the applications of IEA as an automatic way of assessing the content on the basis of Latent Semantic Analysis. The authors explain how IEA's scoring features are combined to score writing, and they evaluate the performance of the IEA model by comparing how IEA's predicted scores match human scoring. They suggest the computation of IEA's reliability by measuring correlation, kappa, weighted kappa, and exact and adjacent agreement, and demonstrate through examples how IEA performs in evaluating the overall quality of an essay, the individual traits of writing, and the content features in short responses. They conclude that IEA provides a means to incorporate accurate scoring.

In Chapter 6, Schultz introduces the theoretical and conceptual bases for IntelliMetric modeling by examining its text features and key principles. He describes the specific process that IntelliMetric undergoes for scoring essays and displays an example of the application in scoring Chinese essays. He concludes that IntelliMetric is accurate in scoring but that the quality of the training set is of utmost importance.

In Chapter 7, Rich, Schneider and D'Brot start with a description of challenges related to a large-scale AEE implementation in West Virginia. They describe how American schools in West Virginia apply AEE technology in summative and formative assessment contexts and explain the engine on which both assessments are based. They continue by analyzing the model of assessment that the West Virginia Education Standard Test is built upon and report three studies on the effect of AEE, in all of which Writing Roadmap™ is employed to examine students' writing performance, with inference that AEE technology could have a positive impact on student writing. They close the chapter by focusing on teacher professional development and future AEE application.

Chapter 8 presents a software tool that enables non-expert learners to use technology to assess texts related to their domains and tasks. Mayfield and Rose begin with an orientation to machine learning and continue with an explanation of models resulting from it. Then they introduce the work-flow of LightSIDE and illustrate how it can be generalized to different tasks in order to help non-expert users achieve a balance between representation complexity and generality. Finally, they report on three studies that focus on essay assessment and the application of machine learning and conclude by discussing the advantages of using machine learning technology, the most important of which is that it enables LightSIDE to adjust to new problems so that the users are not restricted by the basic representations

available on the interface.

Brew and Leacock investigate the use of automated short answer scoring in Chapter 9. They introduce the c-rater engine by explaining how the items are developed, how the model is built, and how the scores are generated. This is followed by a description and explanation of the methods for measuring and assessing the effectiveness of automated scoring. They list some problems of using automated scoring for short answers and suggest solutions to the problems. In the conclusion, they show evidence that automated scoring can benefit short answers, but rather than just giving a single score, it is essential that the programs offer useful and informative feedback.

The next five chapters are concerned with warrants and justifications of AEE. In chapter 10, Williamson provides a framework for validity argumentation of AES. He introduces the validity theory and Toulmin's reasoning as basis for evaluating the strength of validity argument. He then elaborates the elements for argument to explain how Toulmin's model can be applied to human scoring of essays, which in turn provides a contrast for the use of this model for AES. He then considers the similarities and distinctions of human scoring and AES in construct, consistency, and interrelatedness. The chapter closes with implications of the use of AES and directions for improving its technology.

Chapter 11 is concerned with the development of a plan to support the use of AES. Attali starts with the analysis of the validity of features extracted from texts to measure the quality of an essay. Then he turns to the issue of how to combine features to determine the essay score. After that, he discusses the reliability of machine scores by examining the precision of machine scoring across prompts and the coefficients between human and machine scoring. He concludes with suggestions for combining human and machine scores and for how to measure the effects of using AES on student writing.

Chapter 12 is an overview of the methods used in scaling and norming for AES. Koskey and Shermis compare holistic and analytic rubrics and explain the scales for rating the quality of writing samples and then review the methods for forming scales in AEE, including the common standard-setting methods and differential item functioning methods. Throughout this chapter, they raise concerns for validity issues and recommend future efforts to make meaningful scores, both in general and in AES in particular.

In Chapter 13, Bridgeman starts with a summary of the procedures for monitoring and evaluating the quality of human scoring. He explores factors that lead to problematic human scoring and ways to minimize their impact. He also discusses the gold standard for the development and evaluation of machine scoring engines by focusing on how much AEE can imitate human scoring. He concludes that the relationship between human scoring and AES remains a key consideration in the future.

In Chapter 14, Lottridge, Schulz and Mitzel firstly review the literature of problems of human scoring for the purpose of presenting AES as an approach for monitoring human scoring. They then introduce an automated scoring engine named CRASE and report three studies on the identification of human rater bias, the establishment of a criterion for identifying it, and the prevention of the bias through the use of CHASE. They conclude that AES is a significant method for the monitoring of human rater performance.

Chapters 15 to 18 include a number of articles that deal with the analysis of relevant aspects that reflect writing constructs. Chapter 15 is concerned with grammatical error detection in AES and Gamon, Chodorow, Leacock and Tetreault explain the meaning of grammatical error by comparing error categories. Then, they contrast the grammar-based and statistically-based techniques in grammatical error detection. They also give an overview of evaluation issues that include metrics, data, and methods. They finally discuss the practice of four different AES engines (CRASE^{TM}, CTB's AEE, PEG, and e-rater and Criterions systems) and analyze how grammatical feedback affects student writing.

Chapter 16 begins with a discussion of factors that determine the coherence quality of texts. Burstein, Tetreault, Chodorow, Blanchard and Andreyev investigate how to characterize discourse coherence and how linguistic features can be modeled to build an essay evaluating system. In concluding, they

emphasize the importance of including discourse coherence quality ratings in automated scoring and offering explicit feedback about discourse coherence quality in essay evaluation systems.

Another perspective for essay evaluation, sentiment analysis for evaluating argumentation, is addressed in Chapter 17. Burstein, Beigman-Klebranov, Madnani and Faulkner give an overview of related literature on lexicon building. After describing the process of how to create a family of subjectivity lexicons, the authors develop and evaluate an automated sentiment analysis system for identifying the polarity of opinions and the intensity of polarity in students' essays. Their conclusion projects that the system will not only help identify sentiment and polarity in summarization tasks, but will also be applied to different essay modes such as product and movie reviews, and political and newspaper essays.

While most AES frameworks are modeled on the basis of human scoring, in Chapter 18 Deane explores a different approach by introducing a general cognitive framework. He gives an overview of a cognitively based assessment framework and suggests several of its implications for AES, primarily in relation to the selection of features in AES applications and the creation of a criterion involving multiple sources of information. He concludes that the cognitive framework provides a possibility to cover a more comprehensive portion of the writing construct and a new method to support assessment and instruction through the use of AES engines.

Chapter 19 is about a study of the comparison of nine AEE engines. Shermis and Hamner evaluate the performance of these engines on the basis of a set of standard measures: distributional differences, agreement, and agreement delta. Their results show that the overall performance of AEE meets or exceeds that of human raters, but that there is some difference in the performance of each scoring engine. The authors conclude by advocating for a link between public competitors and commercial vendors to provide the best product for essay assessment.

In the final chapter, Hakuta discusses the Common Core State Standards Initiative (CCSSI) and its linguistic challenges and opportunities. The author describes details of CCSSI and points out how it influences major shifts of English Language Arts, Math, and Science in the Standards. The macro-level shifts inherent in CCSSI are shown in Understanding Language Initiative, which emphasizes text-based evidence for argumentation in English Language Arts, the language of reasoning for understanding mathematical practices, and the collaboration of content-area teachers to help students with complex subject-matter texts in science and technology. He concludes that CCSSI has generated new opportunities for scientists and language educators to collaborate, which is a running theme throughout the volume.

The *Handbook of automated essay evaluation* gives a comprehensive illustration of major concerns in the field of AEE. It introduces the theories on the basis of which AEE models are established, demonstrates the practice of AEE applications by both first and second language learners, explains the features and functions of various AEE systems, examines the writing constructs that underlie AEE systems, analyzes the reliability and validity of AEE engines, and predicts the direction for future development of AEE.

What is commendable in this book is that the editors help readers understand the issues better by co-referencing chapters for further exploration. Such an endeavor enables readers to have an overall grasp of the knowledge by relating across many different perspectives. One limitation of the book is that there is no clear division of thematic parts within the collection of articles. Classification into sections could help readers get a clearer picture of the overarching organization of the book. I, therefore, suggest in a future edition of the handbook that chapters be categorized into six parts: introduction, AEE and the teaching of writing, AEE systems, AEE reliability and validity, aspects about writing constructs in AEE, and overall issues of AEE. Despite this limitation, it is undeniable that the book will be of great benefit to those interested in the topic and will serve as a powerful driving force in the development of AEE in the future.

ABOUT THE AUTHOR

Dr. Li Zhang is an associate professor in the School of Foreign Languages in Shanghai Jiao Tong University, China. Her research interests include computer-aided language learning, writing in CALL environments, and writing and communication for academic purposes.

E-mail: zhangli@sjtu.edu.cn

REFERENCES

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48*, 238–243.

Shermis, M. D., & Burstein, J. (2003). Introduction. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. xiii–xvi). Mahwah, NJ: Lawrence Erlbaum.

Waige, S. C. (2013). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 36–54). New York, NY: Routledge.