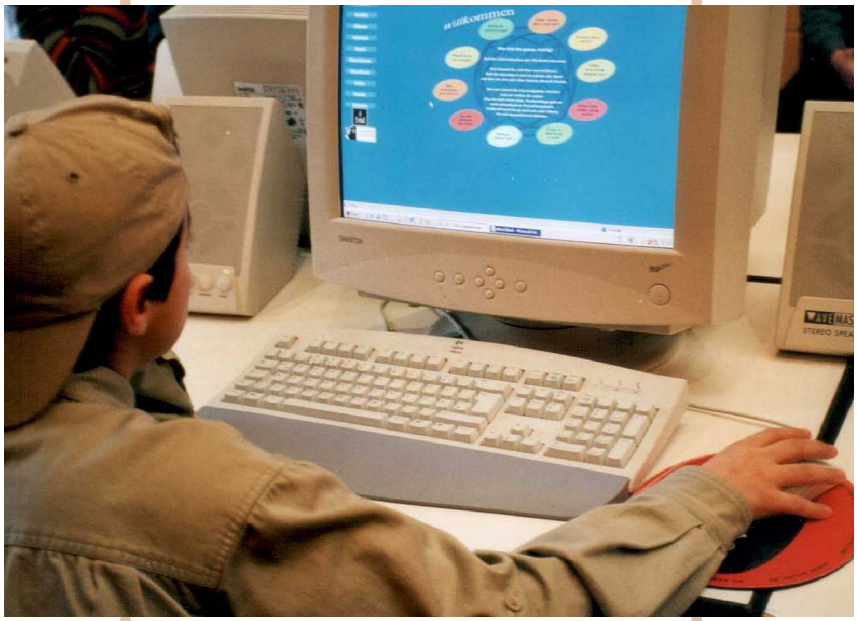
TASA INSTITUTE



AUTOMATED ESSAY SCORING
A LITERATURE REVIEW

SUSAN M. PHILLIPS

SOCIETY FOR THE ADVANCEMENT OF Excellence in Education

Automated Essay Scoring: A Literature Review

Susan M. Phillips

SOCIETY FOR THE ADVANCEMENT OF Excellence in Education

Library and Archives Canada Cataloguing in Publication

Phillips, Susan, 1953-
Automated essay scoring : a literature review / Susan Phillips.
(SAEE research series ; #30)
Includes bibliographical references.
ISBN 978-0-9737755-7-0

Cataloguing details
1. Educational tests and measurements—Computer programs.
2. Grading and marking (Students)—Computer programs.
I. Society for the Advancement of Excellence in Education. II. TASA Institute. III. Title. IV. Series.
LB3060.5.P48 2007 371.26'0285 C2006-906844-5

SOCIETY FOR THE ADVANCEMENT OF EXCELLENCE IN EDUCATION (SAEE)

SAEE is an independent non-profit education research agency founded in 1996. The mission of the Society is to encourage excellence in public education through the provision of research to guide policy and practice. SAEE has commissioned 30 studies to date and has a particular interest in examining innovations that may improve learning outcomes for less-advantaged students. As a registered Canadian charity, SAEE provides official tax receipts for donations to its research.

For additional copies of this report, please contact:

Society for the Advancement of Excellence in Education (SAEE)
225-1889 Springfield Road, Kelowna, BC. V1Y 5V5
Tel. 250-717-1163
info@saee.ca
www.saee.ca

ACKNOWLEDGEMENTS

I would like to express appreciation to the Max Bell Foundation and the Society for the Advancement of Excellence in Education (SAEE) for sponsoring the Technology Assisted Student Assessment Research Institute and providing the research grant in support of this publication. A special thanks goes to Helen Raham, Research Director of SAEE, and Jim Gaskill, Director of TASA.

ABOUT THE AUTHOR

Dr. Susan M. Phillips is an educator with over twenty-five years of experience in the public school system. She obtained her B.Ed. in Secondary Education and M.Ed. in School Counselling from the University of Victoria and her Ed.D. in Educational Leadership from Brigham Young University. She has worked in teaching, counseling, consultant and administrative positions in schools and at the district level as well as being a university sessional lecturer. Her experience encompasses many different types of school organizations and programs including those targeting at-risk adolescents and adult learners.

ABOUT THE TASA INSTITUTE

The TASA Institute is a research initiative of the Society for the Advancement of Excellence in Education (SAEE). Established in 2004 with a research grant from Max Bell Foundation, the mission of TASA Institute is to study and advance knowledge in the development and application of assessment technology in the Canadian public education system.

The purposes of the Institute are:

1. To document trends, leading-edge prototypes, evidence regarding their effectiveness, best practice, and implications for policy in the field of technology-delivered student assessment.
2. To develop a next-generation assessment toolset and process, leveraging the considerable strengths of computer and online technologies.
3. To collaborate with Ministries of Education, school districts, testing agencies and international researchers in the piloting and evaluation of computer assisted assessment models.
4. To serve as a clearinghouse for research and provide a source of expertise to schools, districts, and ministries/departments of education on the design, implementation, and use of computer based assessment.

For more information, see: http://www.tasainstitute.com

EXECUTIVE SUMMARY

Automated Essay Scoring (AES) is a relatively new field that draws upon the diverse disciplines of writing instruction, computational linguistics, and computer science. The purpose of this literature review is to communicate a balanced picture of the state of AES research and its implications for K–12 schools in Canada. It will be of interest to practitioners, developers of assessment technology and educational policy makers.

Susan M. Phillips provides a scan of the most recent literature on this topic which encompasses the variety of AES models, practical issues, diverse perspectives, and directions for future research. This review lays the foundation for the thoughtful use of AES in K–12 schools.

Chapter one describes the state of the literature in the field of AES, the range of stakeholders involved, and the research limitations inherent due to the proprietary nature of many of the AES systems. In chapter two, a history of the development and use of AES is discussed as well as evidence from research indicating advantages and disadvantages related to its use. Chapter three provides an overview of the analytic tools and several of the AES systems available in North America. It also indicates the unique demands that formative and summative assessments make on AES engines. Chapter four presents research on the accuracy and validity of AES systems with respect to writing assessments. Additionally, it provides a background of research done on comparing agreement rates between AES systems and human raters. Chapter five points to key considerations related to K–12 pedagogy with respect to AES, especially in the area of classroom-based formative assessment. This chapter also examines the potential for the use of AES with specific populations such as ESL, Online, and special needs students. Key findings and implications are summarized in chapter six followed by eight recommendations for future research in the areas of pedagogy, technology development and educational policy.

Whether it is viewed as a promising tool to improve the potential of student writing or a threat that removes the teacher from the evaluation process, AES is an issue on the leading edge of assessment for K–12 in Canada.

LIST OF TABLES AND FIGURE

TABLE OF CONTENTS

1. Introduction

As technology becomes an integral component of what and how students learn, its use as an essential tool for student assessment is inevitable.[1]

PURPOSE OF LITERATURE REVIEW

Automated Essay Scoring (AES) in its current formats and applications is a relatively new field that elicits passionate responses from both its supporters and detractors, while still under the radar for the great majority of educators in elementary and secondary schools. Depending on one's perspective, AES is thought to be a promising innovation or a serious threat to the learning of writing. In this review of the literature many different viewpoints, claims, and issues will be brought forward in the hope of stimulating research, policy and practice that will best serve the needs of learners and facilitate the work of teachers.

The purpose of this review is to: describe the technical state of the art in the field of scoring or grading of essays using computer technology; outline past and emergent practices and their pedagogical implications; summarize existing research; direct readers to more specific reference materials; provide a framework/blueprint for thinking about possible risks and potential in this field; and suggest directions for further research and development. Conducted between July and September 2006, this survey is international in scope, but will provide Canadian examples where possible. Research at all levels will be cited, with an emphasis on K–12 when available.

The State of the Literature

While there are numerous journal articles and conference papers related to AES[2] , there appear to be only two books in print that focus on AES. Both Shermis' and Burstein's (eds.) *Automated Essay Scoring: A Cross-Disciplinary Perspective* (2003)[3] and Ericsson's and Haswell's (eds.) *Machine Scoring of Student Essays* (2006)[4] are collections of essays addressing a range of relevant issues as

1 Taylor, A. R. (2006, May). *A Future in the Process of Arrival: Using Computer Technologies for the Assessment of Student Learning*. 22 p. 9. Retrieved from http://www.tasainstitute.com/029.pdf

2 Rudner (2006, p. 5) notes that in January 2005 one online bibliography by Haswell contained over 175 references to machine scoring of essays

3 Shermis, Mark D. and Burstein, Jill (2003). *Automated Essay Scoring: A Cross-Disciplinary Perspective.* Lawrence Erlbaum Associates: Mahwah, New Jersey.

4 Ericsson, Patricia and Haswell, Richard (2006). *Machine Scoring of Student Essays*. Utah State University Press: Logan, Utah.

perceived by their respective editors. It is interesting to note that there does not appear to be a single author published in both books, although the number of people currently doing research in this field is relatively few. The two volumes seem to represent the two ends of the continuum in the discourse related to AES. The literature in this field is growing monthly, however. The most recent article providing a general overview for those who wish to become familiar with the basics of a variety of AES engines was published by Semire Dikli in August 2006.

The majority of the AES literature has been published since the turn of the century (2000–2006), even though the origins of AES are rooted in the 1960's seminal research of Ellis Page. The dramatic increase in the use of AES in large-scale summative evaluations and its application to formative assessment has occurred in the last few years and is projected to continue to develop at an exponential rate.

Range of Perspectives

The development and application of AES represents the merging of at least two or three distinct disciplines: writing instruction, computational linguistics, and computer science. The computer scientists are interested in the logistics and methodology of how the scoring programs operate, e.g., latent semantic analysis. The linguists are interested in the structures of the corpus that are being examined. The writing or English teachers are concerned with the quality and development of writing skills in their students. The teacher wishes to know how to use technology, in this case AES, to advance their students' learning, both individually and collectively.

Needless to say, the varying philosophies and viewpoints held by different disciplines as to the place and importance of AES, its validity and reliability, its effectiveness and efficiency have caused heated debates in this emerging field. Most of the programs are being marketed commercially although in many situations they are being bought by public funds. Their proprietary process is protected by trademark which makes developers reluctant to permit independent research. The National Council of Teachers of English (NCTE) has adopted a position paper banning machine scored writing. Some educators fear that AES is being instituted only to effect cost efficiencies which could result in job cutbacks. Instead of employing multiple human raters for each student paper, AES systems use multiple human raters to score the initial set of papers for the training of the AES engines. Once this set of papers is scored, most AES programs then only use one human rater per student paper to check the accuracy of the AES scoring. Others have concerns that the human raters used in some AES situations are inadequately trained and prepared. These and other realities provide the ingredients for prolonged debate.

RESEARCH LIMITATIONS TO DATE

At this time, little independent comparative research and evaluation has been conducted in this emerging field. Most AES research has been done by personnel associated with one or more of the commercial programs. This has led to inquiry that concentrates more on the effectiveness of a particular engine or program as opposed to the associated pedagogical issues. There are also questions about the validity and reliability of industry-funded research. Additionally, lack of information and varying AES engine characteristics make it very difficult if almost impossible to make direct comparisons between programs. Other issues will be discussed in the reliability and validity chapter.

TECHNOLOGY ASSISTED STUDENT ASSESSMENT

Assessment practices shape, possibly more than any other factor, what is taught and how it is taught in schools. At the same time, these assessment practices serve as the focus … for a shared societal debate about what we, as a society, think are the core purposes and values of education. If we wish to create an education system that reflects and contributes to the development of our changing world, then we need to ask how we might change assessment practices to achieve this.[5]

Internationally educators are echoing the need to critically and continuously review the relationship of authentic assessment, technology and pedagogy if we wish to improve learning for students by continually enhancing professional practice, in concert with and not in reaction to new developments. Years ago Mesthene (1969) stated that "new technology creates new possibilities for human choice and action but leaves their disposition uncertain. What its effects will be and what ends it will serve are not inherent in the technology, but depend on what men will do with technology."[6] More recently Ripley (2004) states "E-assessment must not simply invent new technologies which recycle our current ineffective practices."[7]

In the United Kingdom Ridgway, McCusker and Pead (2004) in *Literature Review of E-Assessment* state that:

> The issue for e-assessment is not if it will happen, but rather, what, when and how it will happen. E-assessment is a stimulus for rethinking the whole curriculum, as well as all current assessment systems. New educational goals continue to emerge, and the process of critical reflection on what is important to learn, and how this might be assessed authentically, needs to be institutionalized into curriculum planning.[8]

In the United States, a report to the President and Congress of the bipartisan Web-Based Education Commission (Kerrey & Isakson, 2000) reached the following conclusion:

> The question is no longer *if* the Internet can be used to transform learning in new and powerful ways. The Commission has found that it can. Nor is the question *should* we invest the time, the energy, and the money necessary to fulfill its promise in defining and shaping new learning opportunity. The Commission believes that we should (p. 134, *italics in original*).

If acted on, the consequences of this statement for assessment are profound. As online learning becomes more widespread, the substance and format of assessment will need to keep pace. The Commission's report also stated:

---

5 Ridgway, Jim, McCusker, Sean, & Pead, Daniel. (2004). *Literature Review of E-Assessment*. 10 p.1. Retrieved from http://www.futurelab.org.uk/research/lit_reviews.htm#lr10

6 Mesthene, E. G. (1969). Some General Implications of the Research of the Harvard University Program on Technology and Society. *Technology and Culture* 10 (4), p. 492.

7 Ripley, M. (2004). E-Assessment: An Overview. *QCA Keynote Speech.*

8 Ridgway, Jim, McCusker, Sean, and Pead, Daniel (2004), p. 4.

> Perhaps the greatest barrier to innovative teaching is assessment that measures yesterday's learning goals…Too often today's tests measure yesterday's skills with yesterday's testing technologies—paper and pencil (p. 59).[9]

Russo states that "computer-based testing can provide flexibility, instant feedback, individualized assessment and eventually lower costs than traditional paper examinations. Computerized results create opportunities for teaching and assessment to be integrated more than ever before and allow for retesting students, measuring growth and linking assessment to instruction."[10] In Canada, Taylor found that:

> An initial search of the literature revealed widespread use of computer technologies in the delivery of assessments in the business and post-secondary education sectors. For example, numerous electronic testing programs are currently utilized with credentialing examinations for entry into the professions and trades, and in the delivery of admissions tests for universities and colleges. Although the K– 12 education system has made significant strides in recent years, it is still in the process of catching up.[11]

Taylor reports an uneven implementation of computer technologies to assess K–12 learning in most provinces and territories.[12] However, student readiness, i.e., familiarity with mouse, screen and keyboard and Internet access, important indicators of readiness to move towards a technology-based form of assessment, is high as Canadian students have greater access to computers than those in many other developed countries.[13]

Benefits of Assessment Technology

Technology assisted student assessment or assessment technology, like paper and pencil testing, has its own unique advantages and limitations. The generic advantages and limitations of assessment technology are briefly summarized here, with more specifics being discussed in other chapters.

Taylor believes that assessment technology offers benefits beyond the capability of paper and pencil testing environments. The potential advantages can include:

- a closer match between curriculum and instruction by enhancement of item types – simulations, models, sound, etc.
- more extensive use of existing banks of items
- greater precision of measures through capacity to adapt to individual student competency levels
- ability to measure learning outcomes not possible through paper and pencil
- cost savings and increased reliability in marking
- greater access for students through the potential for examination on demand

---

9 Bennett, Randy Elliott. (2001). How the Internet Will Help Large-Scale Assessment Reinvent Itself. *Education Policy Analysis* 9 (5), p. 12.
10 Russo, Alexander. (2002). Mixing Technology and Testing. *School Administrator* 4 (59), p. 7.
11 Taylor, Alan R. (2006), pp. 13-14.
12 Ibid., p. 39.
13 Ibid., pp. 49-50.

- vastly improved turnaround time to provide students with instant personalized feedback and teachers with information for remediation and instruction
- savings in shipping, handling, and printing costs
- increased instructional time by reducing labour-intensive marking activities
- increased student ownership of learning through tools to increase their involvement, control and motivation.[14]

These advantages are supported by Ridgway (2004) and others. Bennett (2001) suggests that other advantages of assessment technology for large-scale testing programs include "transmitting some types of complex constructed responses to human graders, removing the need to transport, house, and feed the graders (Odendahl, 1999; Whalen & Bejar, 1998); and scoring other complex constructed responses automatically, reducing the need for human reading (Burstein et al., 1998; Clauser et al., 1997)."[15]

Limitations of Assessment Technology

While there are many advantages to assessment technology, there are also limitations that are not associated with paper and pencil testing methods. As with any innovation, assessment technology requires educators to adapt and to address issues that they may not have had to consider with earlier testing methods. Taylor states that some limitations specific to assessment technology can include:

- issues related to exam delivery – readiness of school for receipt and delivery: capability of computers, sufficient access for students; including sign in procedures and the orientation of proctors and student candidates
- database design quality
- presentation and delivery software quality
- on-site technical expertise or help desk access to resolve unexpected problems
- cost
- potential reliance on private industry for development of appropriate software and database systems [16]

Bennett states that "the inexorable advance of technology will force fundamental changes in the format and content of assessment . . . efforts will need to go beyond the initial achievement of computerizing traditional multiple-choice tests to create assessments that facilitate learning and instruction in ways that paper measures cannot."[17] AES has the potential to do this.

WHAT IS AUTOMATED ESSAY SCORING?

Shermis and others define AES (Automated Essay Scoring) as the computer technology that evaluates and scores written prose. AES systems can both assist teachers with classroom assessment as well as be used in large-scale high-stakes assessment programs by testing companies, states or school districts.

14 Ibid., p. 10-11.
15 Bennett, Randy Elliott. (2001). How the Internet Will Help Large-Scale Assessment Reinvent Itself. *Education Policy Analysis* 9 (5), p. 7.
16 Taylor, Alan R. (2006), p. 29-31, 45.
17 Bennett, Randy Elliott. (2001), p. 1.

Dikli cites a number of authors (Bereiter, 2003; Burstein, 2003; Chung & O'Neil, 1997; Hamp-Lyons, 2001; Myers, 2003; Page, 2003; Rudner & Gagne, 2001; Rudner & Liang, 2002; Sireci & Rizavi, 1999) in supporting her statement that AES systems are mainly used to help overcome time, cost, reliability, and generalizability issues in writing assessment.[18]

As an emerging field, there is no single generic name for these assessment software programs. Indeed, in researching the literature, the following are some of the common terms encountered: AEG - Automated Essay Grading, AES - Automated Essay Scoring, computerized essay scoring, computer essay grading, computer graded essays, computer-assisted writing assessment, machine scoring of essays, AWE - Automated Writing Evaluation, and essay assessment. Terms used appear to be linked to the year of the publication or to the affiliation or trademarked proprietary program being described or to the writer's perspective. For example McAllister and White ". . . prefer 'computer-assisted writing assessment' because it more accurately reflects the current (and previous) state of this discipline." They propose that "virtually none of the work in computer assisted writing assessment is automatic to the point of being autonomous yet, but rather requires numerous human-computer interactions; thus, computers are *assisting* in the partially automated writing assessment process."[19]

Most of the earlier work has "essay" as part of the title. However, many of the more recent innovations are broadening the concept of essay to also include short constructed response answers or free text. Some companies are marketing formative assessment tools that they claim can improve writing without the necessity of teacher mediation. While some suggest that AWE (Automated Writing Evaluation) may become the most common term for computerized scoring of writing, AES (Automated Essay Scoring) will be used as the generic term in this paper.[20]

ESSAYS AND OTHER CONSTRUCTED RESPONSES

Valenti, Nitko and Cucchiarelli (2003) in *An Overview of Current Research on Automated Essay Grading* state that "Essays are considered by many researchers as the most useful tool to assess learning outcomes, implying the ability to recall, organize and integrate ideas, the ability to express oneself in writing and the ability to supply merely than identify interpretation and application of data. It is in the measurement of such outcomes, corresponding to the evaluation and synthesis levels of the Bloom's (1956) taxonomy that the essay questions serve their most useful purpose."[21] In societies influenced by a tradition of British education, the essay exam is often considered a superior evaluation tool[22] to other types of student assessment, e.g. multiple choice questions.

---

18 Dikli, Semire. (2006). An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning, and Assessment* 5 (1).

19 McAllister, Ken S. and White, Edward M. (2006). Interested Complicities in *Machine Scoring of Student Essays*. Edited by Ericsson, Patricia and Haswell, Richard. Utah State University Press: Logan, Utah, p. 246 (Notes).

20 Warschauer, Mark and Ware, Paige. (2006). Automated Writing Evaluation: Defining the Classroom Research Agenda. *Language Teaching Research* 10 (2).

21 Valenti, Salvatore, Nitko, Anthony J., and Cucchiarelli, Alessandro. (2003). An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education* 2 (Information Technology for Assessing Student Learning Special Issue), p. 319.

22 Bereiter, C. (2003). Foreword in *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Edited by Shermis, Mark D. and Burstein, Jill. Lawrence Erlbaum Associates: Mahwah, New Jersey, pp.viii-ix.

Scalise and Gifford (June 2006) created a table (Table 1.1 — Intermediate Constraint Taxonomy for E-Learning Assessment Questions and Tasks)[23] that organizes twenty-eight item examples into a taxonomy based on the level of constraint in the item/task response format. The types of item/task response formats that this literature review will focus on are concentrated in the least constrained, increasingly complex item types that use fully constructed response formats, i.e. essays. Even though AES is now beginning to be applied to some aspects of constructed short answer scoring in different subject areas the main focus of this literature review is the scoring of essays. Emergent research on the next generation of constructed short answer scoring in areas beyond the traditional essay (e.g., science experiments) could be an area of focus for further research at a later date.

The use of technology for student assessment has seen many advances in the last two decades. While most people believe that computer-based assessment can be effective for multiple-choice, true-false and single word answer items, there are now a great number of AES programs that claim to be able to score essays and open-ended items as effectively as human raters. While many herald the event of such software as a boon to pedagogy, there are others who feel the marketing claims are not substantiated. Within the community of writing teachers and instructors, the task of evaluating writing and providing quality feedback is viewed as a complex process,[24] and many within that community believe writing cannot be evaluated appropriately by computer assessment programs.

Indeed, "In 2004, CCCC published a *Position Statement on Teaching, Learning, and Assessing Writing in Digital Environments*, which outlines the criteria effective educational technologies should exhibit, which include hands-on use of technology, its application to specific realms of a student's career or personal life, engaging students in critical evaluation of information, and encouraging reflective practice. It is interesting to note that while much high-quality scholarship has emerged in response to this position statement, the vast majority has dealt with the issues of technology and instruction, rather than with technology and assessment."[25]

Valenti, Nitko, and Cucchiarelli (2003) state that one of the difficulties of grading essays is the perceived subjectivity of the grading process. The subjective nature of essay assessment can lead to variation in grades awarded by different human assessors, which is perceived by students as a great source of unfairness. They believe this issue can be addressed by the use of automated assessment tools for essay scoring. Automated assessment systems would provide consistency in essay scoring, while enormous cost and time savings could occur if the AES system is shown to grade essays within the range of those awarded by human assessors.[26]

---

23 Scalise, Kathleen and Gifford, Bernard. (2006). Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms. *Journal of Technology, Learning, and Assessment* 4 (6), p. 9.

24 Ware, Paige, Warschauer, Mark, and (in press) (2006). Electronic Feedback and Second Language Writing in *Feedback and second language writing*. Edited by Hyland, K. and Hyland, F. Cambridge University Press: Cambridge, England, p.19.

25 Freitz, Elizabeth. (2006). Book Review: Machine Scoring of Student Essays: Truth and Consequences. *The CEA Forum* 35 (1), p. 1.

26 Valenti, Salvatore, Nitko, Anthony J., and Cucchiarelli, Alessandro. (2003)., p. 319.

Warschauer and Ware (2006) believe there is a role for electronic feedback to allow educators to enhance conventional forms of writing or to allow students to become more autonomous learners through increasingly sophisticated computer-generated feedback software. They also think that future research about these newer forms of electronic literacy will explore questions related to novel forms of writing and new ways of teaching and conducting research which will push the boundaries of the forms and functions of electronic feedback in ways that pose new areas for inquiry.[27]

Rudner says that "though some may still find it controversial, automated essay scoring is now widely accepted as a tool to complement, but not replace, expert human raters."[28] Others have differing opinions on the acceptance of AES in writing assessment.

PREVIEW OF REMAINING CHAPTERS

In the following chapters, readers will be introduced to a variety of AES models, issues related to essay evaluation using AES, research to date, and research that needs to occur in order to advance student learning. Chapter 2 will describe the development of AES and discuss the potential advantages and criticisms of this technology.

Chapter 3 will describe some of the analytical tools and classification systems used in various AES engines. These will include Bayesian text classification, latent semantic analysis and natural language processing. A variety of major AES engines such as e-rater®, Intelligent Essay Assessor, Project Essay Grader™ , IntelliMetric™ and Betsy will be described as will two of the formative evaluation programs, Criterion℠, and My Access!®. The differing demands of formative and summative evaluation on AES engines will be discussed and some characteristics and assumptions of different automated essay scoring systems will be compared. The difficulties of direct comparisons in this field due to the proprietary nature and differing analytical tools will be briefly examined.

Chapter 4 looks at reliability issues as they relate to the reliability between human raters and the various AES engines, but also explores some of the concepts of whether or not a future challenge for AES engines is to be better than human raters. Different classifications of human raters will be discussed as will validity issues such as whether AES search engines actually measure attributes that lead to better writing or whether they only measure attributes that lead to better scores. Chapter 5 details some of the pedagogical issues and controversies that arise from the adoption of AES search engines and outlines practices that may need to change in the future in order to best use this new technology to increase student achievement. In conclusion Chapter 6 will state the key findings and their implications and offer recommendations for researchers, educators and policymakers that arise from the literature search.

---

27 Ware, Paige, Warschauer, Mark, and (in press) (2006). *Electronic Feedback and Second Language Writing*. p. 20.

28 Rudner, L. M., Garcia, V., and Welch, C. (2006). An Evaluation of the IntelliMetricSM Essay Scoring System. *Journal of Technology, Learning, and Assessment* 4 (4), p. 4.

Table 1.1 shows twenty-eight item examples organized into a taxonomy based on the level of constraint in the item/task response format. The most constrained item types, at left in Column 1, use fully selected response formats. The least constrained item types, at right in Column 7, use fully constructed response formats. In between are "intermediate constraint items," which are organized with decreasing degrees of constraint from left to right. There is additional ordering that can be seen "within-type," where innovations tend to become increasingly complex from top to bottom when progressing down each column.

Table 1.1 Intermediate Constraint Taxonomy for E-Learning Assessment Questions and Tasks[29]

Most Constrained — Least Constrained

Fully Selected — *Intermediate Constraint Item Types* — *Fully Constructed*

Less Complex ↓ More Complex

1	2	3	4	5	6	7
Multiple Choice	**Selection/ Identification**	**Reordering/ Rearrangement**	**Substitution/ Correction**	**Completion**	**Construction**	**Presentation/ Portfolio**
1A. *True/False* (Haladyna,199 4c, p. 54)	2A. *Multiple True/False* (Haladyna, 1994c, p.58)	3A. *Matching* (Osterlind, 1998, p.234; Haladyna, 1994c, p.50)	4A. *Interlinear* (Haladyna, 1994c, p.65)	5A. *Single Numerical Constructed (Parshall et al, 2002, p. 87)*	6A. *Open-Ended Multiple Choice* (Haladyna, 1994c, p.49)	7A. *Project* (Bennett, 1993, p.4)
1B. *Alternate Choice* (Haladyna, 1994c, p.53)	2B. *Yes/No with Explanation* (McDonald, 2002, p.110)	3B. *Categorizing* (Bennett, 1993, p.44)	4B. *Sore-Finger* (Haladyna, 1994c, p.67)	5B. *Short-Answer & Sentence Completion* (Osterlind, 1998, p.237)	6B. *Figural Constructed Response* (Parshall et al, 2002, p.87)	7B. *Demonstration, Experiment, Performance* (Bennett, 1993, p.45)
1C. *Conventional Standard Multiple* (Haladyna, 1994c, p.47)	2C. *Multiple Answer* (Parshall et al, 2002, p.2; Haladyna, 1994c, p.60)	3C. *Ranking & Sequencing* (Parshall et al, 2002, p.2)	4C. *Limited Figural Drawing* (Bennett, 1993, p.44)	5C. *Cloze-Procedure* (Osterlind, 1998, p.242)	6C. *Concept Map* (Shavelson, R. J., 2001; Chung & Baker, 1997)	7C. *Discussion, Interview* (Bennett, 1993, p.45)
1D. *Multiple Choice with New Media Distracters* (Parshall et al, 2002, p.87)	2D. *Complex Multiple Choice* (Haladyna, 1994c, p.57)	3D. *Assembling Proof* (Bennett, 1993, p.44)	4D. *Bug/Fault Correction* (Bennett, 1993, p.44)	5D. *Matrix Completion* (Embretson, S, 2002, p. 225)	6D. *Essay* (Page et al, 1995, 561-565) & *Automated Editing (Breland et al, 2001,pp.1-64*	7D. *Diagnosis, Teaching* (Bennett, 1993, p.4)

29 Scalise, Kathleen and Gifford, Bernard. (2006). Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms. *Journal of Technology, Learning, and Assessment* 4 (6), p. 9.

2. Automated Essay Scoring

> ***More than many issues within the field of composition studies, writing assessment evokes strong passions.***[30]

No one denies that the impact of technology has affected many facets of modern life including the process of writing and communication. What people do not agree upon is whether technology can and should be used in the teaching and assessment of writing.

While AES generates strong passions and a number of areas of disagreement among different stakeholders, there are some areas of agreement among educators, software developers and the general public. They include the following assumptions or beliefs:

- writing is multifaceted.
- there is empirical evidence that individuals can get better at writing.
- developing writers require support.
- people learn to write by writing.

In Canada at least five provincial teachers associations are listed as official affiliates of National Council of Teachers of English (NCTE). The Canadian Council of Teachers of English Language Arts (CCTELA) is a member of International Federation for the Teaching of English as is NCTE. The following position paper, *NCTE Beliefs About the Teaching of Writing* will be referenced throughout this literature review as will the *CCCC Position Statement on Teaching, Learning, and Assessing Writing in Digital Environments* (see Appendix A) and *Writing Assessment: A Position Statement* also written by the CCCC (Conference on College Composition and Communication) which is part of the NCTE organization.

The NCTE states that:

> Just as the nature of and expectation for literacy has changed in the past century and a half, so has the nature of writing. Much of that change has been due to technological developments, from pen and paper, to typewriter, to word processor, to networked computer, to design software capable of composing words, images, and sounds. These developments not only expanded the types of texts that writers produce, they also expanded immediate access to a wider variety of readers. With full recognition that writing is an increasingly multifaceted activity, we offer several principles that should guide effective teaching practice.[31]

---

30 Conference on College Composition and Communication (1995). *Writing Assessment: A Position Statement*. p.1. Retrieved from http://www.ncte.org/cccc/resources/positions/123784.htm

31 Writing Study Group of the NCTE Executive Committee (2004). *NCTE Beliefs About the Teaching of Writing*. p.1. Retrieved from http://www.ncte.org/about/over/positions/category/write/118876.htm

Among those principles, it is clear that NCTE does not sanction AES as it states "we oppose the use of machine-scored writing in the assessment of writing"[32] and that "all student writing, including college entrance exams, are evaluated by knowledgeable humans rather than scored by machines."[33]

Advocates of AES cite the possibility of immediate feedback or at least a quicker turnaround time than that for paper and pencil assessment, more consistent feedback and the ability to process more work per student to increase opportunities to improve writing and for students to receive feedback as some of the benefits AES can provide.

Further study of the characteristics of AES is appropriate because writing has always been an important skill for students and professionals and indeed through writing we often assess other forms of knowledge. Given that the technology for AES exists and is being marketed by various companies and individuals with new programs appearing regularly, it is essential that AES be examined by all stakeholders in a comprehensive manner.

BRIEF HISTORY OF THE DEVELOPMENT AND USE OF AES

Burstein states that Educational Testing Service (ETS) has been conducting research in writing assessment since 1947 and administered the Naval Academy English Examination and the Foreign Service Examinations as early as 1948 and the Advanced Placement (AP) essay exam in 1956.[34]

Page (2003) describes the earliest developments in AES systems as originating from the demands of grading writing assignments, either as a teacher or instructor in individual classrooms or large-scale demands such as the hundreds of thousands of essays the College Board was manually grading by 1964. After some initial trials that were funded by the College Board, additional private and federal support was provided which helped develop a program of research at the University of Connecticut. Page describes their early research as successful but notes that full-scale implementation barriers were practical in nature[35], i.e. data input to computers was accomplished through tape and/or 80 column IBM punch cards; processing speeds of computer mainframes were relatively slow; processing was primarily in batch mode and was not very tolerant to unanticipated errors; and what he considered the major concern – access to computers was restricted for the vast majority of students.

The development of Project Essay Grader™ (PEG) went into "sleep mode" until the mid-80s when the advent of microcomputers permitted a number of technological advances which prompted re-examination of the potential for automated essay scoring.

---

32 Ibid., p. 14.

33 Assessment and Testing Study Group of the NCTE Executive Committee, (2004). *Framing Statements on Assessment: Revised Report of the Assessment and Testing Study Group of the NCTE Executive Committee*. p.4. Retrieved from http://www.ncte.org/about/over/positions/category/assess/118875.htm

34 Burstein, J. (2003). The E-rater® Scoring Engine: Automated Essay Scoring with Natural Language Processing in *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Edited by Shermis, Mark D. and Burstein, Jill. Lawrence Erlbaum Associates: Mahwah, New Jersey, p. 113.

35 Page, Ellis Batten (2003). Project Essay Grade: PEG in *Automated Essay Scoring: A Cross-Disciplinary Perspective.* Edited by Shermis, Mark D. and Burstein, Jill. Lawrence Erlbaum Associates: Mahwah, New Jersey, pp.44-45.

The ETS commissioned a blind test of the Praxis essays (used in evaluating applicants for teacher certification) using PEG, which found that PEG predicted human judgments as well or better even than three human judges. Page describes the expected benefits of computer ratings "to be superior to the usual human ratings in a striking number of ways":

- The automated ratings would surpass the accuracy of the usual two judges. (Accuracy is defined as agreeing with the mean of judgments.)
- The essays would be graded much more rapidly, because fewer human reading would be required.
- Machine-readable protocols would be graded more economically, saving 97% of the grading costs.
- Essay results could be described statistically in many different ways, and used to study group differences, yearly trends, teaching methods, and a host of other important policy or research questions. (Such reports from human graded efforts are often time-consuming and costly.)
- For individual accuracy of writing abilities, scores would be much more descriptive than the ordinary ratings results from two human judgments.
- Validity checks could be built-in to address potential biases (computer or human).[36]

In 1988 Landauer and colleagues first conceived the basic statistical model for latent semantic analysis and by 1996 Foltz was using a prototype of what would become Intelligent Essay Assessor to score essays in his own psychology classes. After incorporating as KAT or Knowledge Analysis Technologies, Landauer and Foltz put Intelligent Essay Assessor online. Harcourt Achieve used their services to score General Educational Development (GED) test practice essays; Prentice-Hall scored assignments in textbooks; Florida Gulf Coast University scored essays by visual and performing arts general education students; the US Department of Education developed "auto-tutors"; and US armed services assessed exams during officer training. Pearson Education acquired KAT in 2004.[37]

Haswell (2006) states that the same pattern of development is evident in at least two other instances: (1) at ETS where e-rater® was developed by Jill Burstein and others and first used publicly to score GMAT essays in 2002; and, (2) at Vantage Laboratories where Scott Elliott developed IntelliMetric™ which was first put online in 1998 and emerged as the model for College Board's WritePlacer, the essay grading component of ACCUPLACER in 2003. In February 1999 ETS started to use e-rater® for the scoring of the GMAT Analytical Writing Assessment (AWA). This continued until January 2006 when the IntelliMetric™ essay scoring system was adopted.

The pattern of development that Haswell identifies is that AES emerged during the 1990s from computer linguistic analysis and information retrieval tools. Haswell states writing teachers had never shown interest in or had abandoned the use of these tools. These tools included machine translation, automatic summary and index generation, corpora building, vocabulary and syntax and text analysis. He believes that researchers and teachers in disciplines other than writing instruction filled the gap as it wasn't

36 Ibid., p. 46.

37 Haswell, Richard (2006). Automatons and Automated Scoring in *Machine Scoring of Student Essays*. Edited by Ericsson, Patricia and Haswell, Richard. Utah State University Press: Logan, Utah, pp. 61-62.

filled by writing researchers and teachers. Different kinds of software are flourishing and making profits for industry in foreign language labs, ESL labs, job-training labs, officers' training schools, textbook and workbook publishing houses, test-preparation and distance-learning firms, online universities, plagiarism detectors and, as he describes it, the now ubiquitous computer classrooms of the schools.[38]

McAllister and White view the development of computer-assisted writing assessment as a complex evolution driven by the dialectic among researchers, entrepreneurs, and teachers; "the history of computer-assisted writing assessment using a broad perspective that takes into account the roles of computational and linguistics research, the entrepreneurialism that turns such research into branded commodities, the adoption and rejection of these technologies among teachers and administrators, and the reception of computer-assisted writing assessment by the students whose work these technologies process."[39] They believe that AES is at the point where the balance of funding is slowly shifting from the research side to the commercial side and where there is increasing acceptance that computers can prove useful in assessing writing despite the protestations of many teachers and writers.

No matter who is recounting the history of AES, there is agreement that it was developed by people from different disciplines trying to address the issues related to the assessment of writing, that the growth of AES has been exponential especially since the turn of this century, and that the numbers of students assessed and types of uses considered is increasing rapidly.

ADVANTAGES OF AES

Hearst (2000) states that computer use to increase understanding of the textual features and cognitive skills involved in the creation and in the comprehension of written text is beneficial to the educational community. These benefits include the development of more effective instructional materials for improving reading, writing and other communication abilities and the development of more effective technologies such as search engines and question answering systems which provide universal access to electronic information.[40] More specifically, Dikli comments that AES systems are primarily used to address time, cost, reliability, and generalizability issues in writing assessment (Bereiter, 2003; Burstein, 2003; Chung & O'Neil, 1997; Hamp-Lyons, 2001; Myers, 2003; Page, 2003; Rudner & Gagne, 2001; Rudner & Liang, 2002; Sireci & Rizavi, 1999).[41]

---

38 Ibid., p. 63.

39 McAllister, Ken S. and White, Edward M. (2006). Interested Complicities. p. 9.

40 Valenti, Salvatore, Nitko, Anthony J., and Cucchiarelli, Alessandro. (03). An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education* 2 (Information Technology for Assessing Student Learning Special Issue), p. 320.

41 Dikli, Semire. (2006). An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning, and Assessment* 5 (1), p. 4.

Assessment Occurring in Same Milieu as Learning

Goldberg, Russell and Cook (2003) conducted meta-analyses of more than 26 studies carried out between 1992 and 2002, which focused upon K-12 students writing with computers versus paper and pencil. They found significant mean effect sizes in favor of computers…

> for quantity of writing (d=.50, n=14) and quality of writing (d=.41, n=15). Studies focused on revision behaviors between these two writing conditions (n=6) revealed mixed results. Others studies collected for the meta-analysis (which did not meet the statistical criteria) . . . indicate that the writing process is more collaborative, iterative, and social in computer classrooms as compared with paper-and-pencil environments. For educational leaders questioning whether computers should be used to help students develop writing skills, the results of the meta-analyses suggest that on average students who use computers when learning to write are not only more engaged and motivated in their writing, but they produce written work that is of greater length and higher quality. [42]

Many students are learning to write in classrooms on computers using word processing programs but are being assessed by paper and pencil means. Assessing students in the technological milieu or context in which they usually write is more educationally sound. This is especially relevant as many of the professions that the students will enter in the future use word processing, communication and other computer applications.

Objectivity and Plagiarism

AES scoring eliminates any human rater subjectivity as the AES program scores every piece of written work for a given prompt in exactly the same way. Unlike human raters it never experiences fatigue or is distracted by extraneous events so reliability is not affected.

Plagiarism appears to be a growing concern at all levels of education and in many countries. In Canada, Julia Christenson Hughes and Donald McCabe found 73% of first-year university students reported instances of serious cheating on written work while in high school.[43] Some AES engines can detect plagiarism, at least among the cohort of papers submitted for assessment in response to a specific prompt. Plagiarism is much easier to detect consistently by machine than by human raters.

Efficiencies

There is a range of efficiencies that AES can affect in terms of marking time, shipping costs, and ease of access to statistical data.

Marking Time at Classroom Level

Among its advocates a major perceived benefit of AES is the reduction of marking time required by

---

42 Goldberg, Amie, Russell, Michael, and Cook, Abagail. (2003). The Effect of Computers on Student Writing: A Meta-analysis of Studies from 1992 to 2002. *Journal of Technology, Learning, and Assessment* 2 (1), p. 1.

43 Sarah Schmidt (2006, September). Most Undergrads Admit to Cheating, Study Finds. *Times Colonist*

teachers in order to provide feedback to students. Naylor and Malcolmson (2001) state that a full-time secondary English teacher in British Columbia spends an average of 11.5 hours a week marking. BC English teachers report three factors in their marking workload: (1) the sheer quantity of marking is excessive, (2) marking itself is highly demanding because it involves reading and assessment of student writing, and (3) the lack of recognition of English marking requirements and the inadequate time provided for marking results in unacceptable levels of fatigue and exhaustion.[44] A study of Alberta high school teachers noted marking consumed 12.84% of their total work time for a week. This data correlated highly with data collected elsewhere in Canada at both provincial and national levels.[45]

Internationally the typical secondary teacher in most countries will have well over 100 students and while numbers vary widely in university classes Warschauer and Ware (2006) state that large class sizes are the norm in many EFL settings. They also claim that many EFL instructors do not have the level of skill and training to provide clear and detailed feedback on writing.[46] Valenti et al. (2003) cite Mason (2002) who states that about 30% of teachers' time in Great Britain is devoted to marking. Mason recommends that if we want to access those resources (1/3 of teacher costs) "then we must find an effective way that teachers will trust, to mark essays and short text responses."[47]

Marking Time in Large-Scale Assessments

Another perceived benefit of AES is the reduction of costs associated with the scoring of essays in large-scale assessments. With the potential for scoring essays at speeds of 2½ minutes per paper, immediate electronic transmission of prompts, student work and assessment results, and the reduction of number of human raters per paper, costs can be greatly decreased. Large numbers of essays can be scored within short time periods: the state of Virginia scored 44,000 papers in 2005; Texas scored 1.7 million.[48]

The high speed of marking reduces the number of human marker work days required and therefore the wage costs dramatically. Some assessment programs have even further cut their human marker costs by employing scorers instead of releasing educators from the classroom to be involved in the scoring sessions. Often the scorers are paid an hourly wage lower than that typical for teachers. In this scenario there is also no cost for substitute teachers, travel and per diem costs. Not everyone agrees with this method as it is felt by some that opportunities for professional development are lost if teachers do not score student work.

---

44 Naylor, C. & Malcolmson, J. (2001, September). BCTF Research Report 2001-WLC-02 *"I Love Teaching English, but...." A Study of the Workload of English Teachers in B.C. Secondary Grades.* RT01-0036 pp.16-17. Retrieved from www.bctf.ca/ResearchReports/2001wlc02

45 Alberta Teachers' Association (2002). *Alberta Teachers' Association: Ongoing Issues - Hours of Instruction*. p. 1-2. Retrieved from http://www.teacher.ab.ca/Issues+In+Education/Ongoing+Issues/Hours+of+Instruction.htm

46 Warschauer, Mark and Ware, Paige. (2006). Automated Writing Evaluation: Defining the Classroom Research Agenda. *Language Teaching Research* 10 (2), p. 2.

47 Valenti, Salvatore, Nitko, Anthony J., and Cucchiarelli, Alessandro. (2003). p. 320.

48 Berard, Y. & Cromer Brock, K. (2006, July). Who's Keeping Score? Grading the TAKS Essay. *Star-Telegram* Retrieved from http://www.dfw.com/mld/dfw/community/15105085.htm

Reduction or Elimination of Shipping Costs

Even further cost reductions are realized when there are little or no shipping costs or delays if assessments are able to be completed on-line. This does presume however, that there is sufficient access to computers for all students involved in the assessment.

Ease of Access to Statistical Data

Electronic scoring and transmission of assessment results allows for greater ease in sharing and manipulation of statistical data so that educators and administrators can use the information for a variety of purposes. Trends, anomalies, and multi-year tracking are among the ways the data can be useful for informing instruction.

CRITICISMS OF AES

AES, similar to all testing, is criticized by some for as Page (2003) states that "all important testing does a job that is inherently unpopular. It differentiates among individuals, and often ties these differences to major decisions—admission to selective programs, professional advancement, licensing, and certification. Just as multiple-choice testing still has many critics, we can expect that computerized grading of essays will continue to be problematic for some."[49] Dikli categorizes the criticisms as lack of human interaction, vulnerability to cheating, and the need for a large corpus of sample essays to train the AES engines.[50] Page classifies the objections as humanist objections, defense objections and construct objections.

Humanistic Objections

The *CCCC Position Statement on Teaching, Learning, and Assessing Writing in Digital Environments* (2004) states that "because all writing is social, all writing should have human readers, regardless of the purpose of the writing." Their reasons for this statement include: writing to a machine violates the essentially social nature of writing; they do not know the criteria and therefore the bias by which the computer scores the writing; and if the student's first writing experience at an institution is writing to a machine then this may send a message that writing is not valued as human communication therefore reducing the validity of the assessment.[51]

Page flatly rejects the humanist objections, i.e. that human knowledge and background wisdom are necessary for certain tasks as computers only do what they are programmed to do. He states that if we were to do the "Turing Test" of passing essays under seven doors to be scored, we would quickly identify the computer as it produces the scores that agree best with the other six human judges.

---

49 Page, Ellis Batten (2003). Project Essay Grade: PEG. p. 51.

50 Dikli, Semire. (2006). p. 4.

51 Conference on College Composition and Communication (2004). *CCCC Position Statement on Teaching, Learning, and Assessing Writing in Digital Environments*. Retrieved from http://www.ncte.org/cccc/resources/positions/123773.htm

Defensive Objections

Defensive objections are questions about the assumptions of the essay environment. There is concern that in the real world there is a need to defend against students who generate essays under "bad faith" conditions. Page states there are a number of strategies that can be undertaken to protect against such possibilities and that most AES systems already have subroutines to alert the program about this type of writing, e.g., the use of common vulgarities. He continues that the setting aside of such essays is already common in large-scale human judge assessments when the essays are determined as off-the-subject or unreadable, or when the differences between different judge's ratings are unacceptably large. He believes the overall defensive objection is moot as the majority of AES applications require at least one parallel human reading for high stakes situations.[52]

Construct Objections

Construct objections focus on whether the program measures variables that are important to writing quality. Page responds that how we judge the qualifications of human raters (the ones that are asked return year after year) is by their correlations with other human judges. In that regard AES programs like PEG have a good correlation with other human raters and therefore would be asked to return year after year as they would appear to be measuring variables that are considered important to writing quality by other human raters. He states there is always going to be the question about whether the human raters that were asked not to return, because their correlations with other human raters were so low, were maybe "more qualified" than those human raters that were retained but this is a question for all writing assessment not just AES models.

Financial and Logistical Considerations

As already stated in the introductory chapter under the general limitations of assessment technology, there may be a number of financial and logistical issues related to computer-assisted scoring that might mitigate against using AES for essay scoring. Some of these considerations may be of greater importance in large summative evaluations than in smaller formative evaluations. For example, it may be easier in a formative classroom evaluation setting for all students in the cohort (which may be a single student, small group or class) to have access to computers and the Internet simultaneously. In large-scale assessments where simultaneous access by all members of the cohort (which may be as large as all students in a specific grade in a province or state) is required there may be a problem of computer access. As well in large-scale assessments requiring simultaneous access, if computers are arranged in small clusters, i.e. classrooms or computer labs, there may be hidden costs of providing additional supervisors or exam proctors. In traditional large-scale paper-and-pencil assessments, students are often gathered in large open areas such as gymnasiums where they can be supervised by a minimum number of exam proctors. If the same assessment is being delivered in computer labs, a greater number of proctors will be required as there will be fewer student stations per room.

---

52 Page, Ellis Batten (2003). Project Essay Grade: PEG. p. 52.

Technical expertise needs to be available to resolve unexpected problems, and this may also be an additional cost that would not be present in paper and pencil assessments. The actual costs for the software and scoring services will vary depending upon the numbers of students being assessed and who is responsible for the financial costs of that assessment.

At this time in Canada, there is no systematic cost-benefit analysis of AES as it does not appear AES is implemented in K–12 settings yet. When AES is piloted in Canada, care should be taken to do a cost-benefit analysis in a systematic and comparative way for both formative and summative evaluation situations. As stated in the recommendations section of this paper, a proactive role needs to be taken by professionals from the Canadian K–12 education teaching, assessment and policymaking communities to ensure that the adoption of AES, if it occurs, is as educationally effective and efficient as possible.

Other Objections

Condon (2006) states that even if all the claims in favor of machine scoring are true, i.e. that machines can score similar to what human readers can do, there still will be corresponding losses if machine scoring is implemented. He lists these losses as: (1) writing teachers lose control of the writing construct as what they assess is dictated by an outside agency or by the capacity of the rating machine resulting in a loss of local agency; (2) writing samples must be short, preventing raters from taking original approaches to topics; (3) because of time restraints, topics must be far simpler than the topics most teachers use in class. He concludes that these losses limit the face validity of the sample, meaning that only very limited conclusions can be drawn from the sample. He also states that scoring a set of timed writings for placement (deciding which level or course of writing instruction is appropriate for a student) and again for critical thinking, results in scores that show a negative correlation. Another major loss he predicts is that of conversations about writing, about writing standards, and about judgments of quality. In discussing how AES curtails conversations around writing, Condon states there is a loss of professional development, a loss of exchange of information about students, and a loss of a rich knowledge set from instruction into assessment and from assessment back into instruction.

Broad (2006) counters AES supporters' claim that the time writing instructors spend on assessment is time taken away from teaching. For him, the crucial question is whether students, teachers and the general public should accept this view and endorse it by buying products that he believes help to separate teaching from assessment. Broad acknowledges that "evaluation of writing holds an undeniably murky and ambiguous place in the hearts of most writing teachers"[53] since many openly dread evaluating their students' writing due to the tremendous time and effort and often dubious pedagogical benefits. He believes this does not have to be the only scenario because many teachers "do some of our highest quality teaching responding to and evaluating our students' writing" and that "teachers of writing need to reclaim assessment as a crucial, powerful, and rewarding part of the process of teaching and learning writing."[54]

---

53 Broad, Bob (2006). More Work for Teacher? Possible Futures of Teaching Writing in the Age of Computerized Writing Assessment in *Machine Scoring of Student Essays*. Edited by Ericsson, Patricia and Haswell, Richard. Utah State University Press: Logan, Utah, p. 224.

54 Ibid., p. 224.

Haswell notes that scoring machines promise three things for your money: efficiency, objectivity, and freedom from drudgery—three goals that writing teachers have been trying to achieve in their own practices by way of evaluation for over a century. However, he sets forth an agenda that he believes is best for writing assessment and that it may or may not include AES as it is currently implemented. He does second the call of Williamson (2004, p. 1000) for the writing discipline "to study automated assessment in order to explicate the potential value for teaching and learning, as well as the potential harm." [55]

FUTURE CHALLENGES

Bereiter[56] states that *Automated Essay Scoring: A Cross Disciplinary Perspective* (2003) is a coming of age book as he believes AES is still a young science, but it is ready to venture forward. He believes that the early years were spent proving that the computer can do as well as human raters in scoring a large number of compositions produced under similar conditions in response to the same "prompt." Even during the 1960s in Ellis Page's research, a computer could yield scores that agree with human raters as well or better than the human raters agree with each other. He states that performance gains since then have been incremental even though the algorithms and technology have become increasingly sophisticated.

A frontier in AES development that he believes needs to be addressed is for AES scoring to do better than human raters. Human raters are not perfect as they are susceptible to quirks, biases and halo effects when scoring as well as other issues (e.g., fatigue, training, etc.). Research up to this time has taken the criteria of human raters as the standard but he believes that in the future we may instead try to match some external criteria.

Human raters also are not necessarily equal in their scoring ability: expert readers/raters may be better than other human raters. The use of expert raters may also cause some concerns as he believes that the correlation between ratings in style and content is probably higher than it deserves to be. He relates an editor's experience of peer review of manuscripts submitted to scholarly journals. The editor states that bad writing was never given as the stated reason for rejection of an article even though the expert peer reviewers never recommended a badly written article for publication. Another approach to develop enhanced criteria may be to use expert writers as raters rather than expert readers as raters. However, Sarah Friedman, in her 1980s research found that holistic ratings by human raters did not necessarily score professionally written essays particularly higher compared to those written by students.

Bereiter proposes a number of challenges for AES. The first is to transform AES into a learning tool, i.e., give students the opportunity to score practice essays or drafts to improve their scores without teacher intervention though some teachers may want to provide guidance. However, a basic problem in the "no intervention" or non-teacher mediated situation is that learners are generally not competent to judge the validity of the writing advice provided by the scoring program.
Another challenge, already taken up by the Intelligent Essay Assessor Group, is to score essay

55 Haswell, Richard (2006). Automatons and Automated Scoring. pp. 74-77.
56 Bereiter, C. (2003). p. 7.

examinations as distinguished from essay writing tests. The essay examination mostly tests content knowledge and understanding rather than the composition skills. In societies influenced by British education, the essay exam is often considered a superior evaluation tool. However, when used for mass testing, they attract most of the drawbacks attributed to objective tests, e.g., raters are provided a checklist of points to score. He argues that scoring in a mechanical way is more appropriate for a computer program than for a fatigue prone human.

Bereiter hypothesizes that AES could help to break the "mutual stranglehold that exists between tests and curricula, where the curriculum is constrained by what is on the tests and the tests are derived from what is conventionally taught."[57] Bereiter notes that AES should be able to yield usable results even if students are not all responding to the same prompt. It should also be sensitive to indications of depth of understanding and not merely to the quantity of facts.

According to Shermis and Burstein, the editors of *Automated Essay Scoring: A Cross-Disciplinary Perspective*, a primary challenge is to develop AES and evaluation capabilities so that they are consistent with the needs of educators and their students in early education, secondary and higher education internationally[58] . Development of AES technology has been met with many concerns and questions. Shermis and Burstein state that teachers want to know: how the technology works, how it can supplement classroom instruction, and whether or not it addresses relevant issues that can improve their students' writing. Researchers in educational measurement question the reliability of the technology; computer scientists are interested in the technology methods and capabilities.

Automated Essay Scoring: A Cross-Disciplinary Perspective presents the evolution and the current developments of automated essay scoring—an evaluation technology across the disciplines of teaching pedagogy, educational measurement, cognitive science, and computational linguistics. It examines: "(a) how automated essay scoring and evaluation can be used as a supplement to writing assessment and instruction, (b), several approaches to automated essay scoring systems, (c) measurement studies that examine the reliability of automated analysis of writing, and (d) state-of-the-art essay evaluation technologies."[59]

In contrast Ericsson and Haswell, editors of *Machine Scoring of Student Essays*, state that "the analysis and scoring a student essays by computer—the history, the mechanisms, the theory, and the educational consequences—is the topic of this collection of essays."[60] They assert that "it is an understatement to say the topic is growing in importance at all levels of the educational enterprise, and perspective on it has been, up to this point, dominated almost exclusively by the commercial

---

57 Bereiter, C. (2003). pp. viii-ix.

58 Shermis, Mark D. and Burstein, Jill (2003). *Preface in Automated Essay Scoring: A Cross-Disciplinary Perspective*. Edited by Shermis, Mark D. and Burstein, Jill. Lawrence Erlbaum Associates: Mahwah, New Jersey.

59 Ibid., p. xi.

60 Ericsson, Patricia and Haswell, Richard (2006). Introduction in *Machine Scoring of Student Essays*. Edited by Ericsson, Patricia and Haswell, Richard. Utah State University Press: Logan, Utah, p. 1.

purveyors of the product."[61]

Despite the differing viewpoints, it is clear that AES is a topic of growing importance at all levels of education.

61 Ibid., p. 1.

3. Automated Essay Scoring Systems

This chapter provides an overview of the analytic tools and several of the major AES systems currently in use in North America as well as the differing demands placed upon AES systems dependent upon whether AES is to be used for formative or summative evaluation. Some of the characteristics of selected AES systems are compared; however, a thorough direct scientific comparison is not available for reasons outlined at the end of this chapter.

ANALYTIC TOOLS

There are several different techniques used by the various AES programs. The most common models are Bayesian Text Classification, Latent Semantic Analysis (LSA) and Natural Language Processing (NLP). There are many descriptions of these techniques. However, many of the workings of these techniques especially as they relate to AES are private and covered by trademark protection. As there are entire papers devoted to some of these techniques, a brief description of each of the tools and AES systems will be provided in this review. Readers interested in more detailed descriptions should refer to the bibliography. Readers who wish a quick overview of each of these systems but in more detail than provided here are referred to *An Overview of Automated Scoring of Essays* (Dikli, August 2006)[62] Dikli discusses the following systems and current issues regarding their use in classrooms and in standardized testing: Project Essay Grader™ (PEG), Intelligent Essay Assessor™ (IEA), Criterion℠, e-rater®, IntelliMetric™, MY Access® and BETSY.

Another excellent overview is provided in *An Overview on Current Research on Automated Essay Grading* (Valenti, Neri and Cucchiarelli, 2003). Valenti et al. discuss the general structure and claimed performances of ten systems: Project Essay Grade (PEG), Intelligent Essay Assessor™ (IEA), Educational Testing Service I, Electronic Essay Rater (e-rater®), C-Rater, BETSY, Intelligent Essay Marking System, SEAR, Paperless School free text Marking Engine and Automark. The authors then attempt to compare the performances of the described systems as they existed in 2003."[63] These two references provide much of the material for this chapter. Information from other authors, many of

62 Dikli, Semire. (2006). An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning, and Assessment* 5 (1), p. 4.

63 Valenti, Salvatore, Nitko, Anthony J., and Cucchiarelli, Alessandro. (03). An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education* 2 (Information Technology for Assessing Student Learning Special Issue), p. 319.

whom are associated with a specific proprietary program, is also cited.

Despite the variety of approaches used in AES engines Rudner, Garcia and Welch state that the basic procedure is similar in all cases. "A relatively large set of pre-scored essays responding to one prompt is used to develop or calibrate a scoring model for that prompt. Once calibrated, the model is applied as a scoring tool. Models are then typically validated by applying them to a second, but independent, set of pre-scored items."[64]

Bayesian Text Classification

Bayesian pertains to statistical methods that regard parameters of a population as random variables having known probability distributions and are based on Thomas Bayes' probability theorem involving prior knowledge and accumulated experience. The two common Bayesian models used in text classification are the Multivariate Bernoulli Model and the Multinominal Model. Deferring to Dikli (2006) who explains:

> While the former views each essay as a special case of calibrated features, the latter views each essay as a sample of calibrated features. In the Bernoulli model, the conditional probability of presence of a specific feature is estimated by the proportion of essays within each category that include the feature. In Multinomial model, on the other hand, the probability of each score for a given essay is computed as the product of the probabilities of the features included in the essay. (BETSY, n.d.; Rudner & Liang, 2002). To summarize, the Bernoulli model investigates whether a specific feature exists in an essay or not, whereas the Multinominal model checks the multiple use of a specific feature in an essay (Rudner & Liang, 2002). The Bernoulli model computes relatively slowly compared to the Multinominal model (BETSY, n.d.).
>
> The Bayesian approach includes key concepts such as *stemming, stop words,* and *feature selection.* Stemming denotes the process of eliminating suffixes to get stems. For example, obtaining "educ" as a stem for educate, education, educates, educational, and educated. Stop words refer to various articles, pronouns, adjectives, and prepositions. Search engines do not list these types of words because they can cause large number of irrelevant results. One approach to feature selection is the reduction in *entropy.* By minimizing entropy, it is possible to pick the items with maximum potential information gain (Rudner & Liang, 2002).[65]

Readers who wish a more comprehensive discussion of the application of hierarchical statistical methods using Bayesian methodology for the scoring of essays by multiple readers are referred to Ponisciak and Johnson (2003).[66] They examined the relationship among scores across raters and categories of six raters on a 1 to 6 scale for 1200 essays. They concluded that the relationship between

64 Rudner, L. M., Garcia, V., and Welch, C. (2006). An Evaluation of the IntelliMetricSM Essay Scoring System. *Journal of Technology, Learning, and Assessment* 4 (4), p. 5.

65 Dikli, Semire. (2006). p. 20.

66 Ponisciak, Steve and Johnson, Valen (2003). Bayesian Analysis of Essay Grading in *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Edited by Shermis, Mark D. and Burstein, Jill. Lawrence Erlbaum Associates: Mahwah, New Jersey.

category ratings for a given rater is stronger than the relationship between raters for a fixed category.

Latent Semantic Analysis (LSA)

Thomas Landauer says he and colleagues first conceived of the basic statistical model for latent semantic analysis in 1988 and by 1989 it was patented, the start of a path that led to the commercial success of Intelligent Essay Assessor.[67] Foltz (1996) defines Latent Semantic Analysis as "a statistical model of word usage that permits comparisons of the semantic similarity between pieces of textual information." LSA represents documents and their word contents in a large two-dimensional matrix semantic space. Singular Value decomposition (SVD), a matrix algebra technique, is then used to determine the relationships between words and documents and existing relationships are modified to more accurately represent the true significance. Further analysis of these matrices through SVD allows the scoring to occur. "According to this method, the semantic information is determined only through the co-occurrence of words in a large corpus of texts."[68]

Natural Language Processing (NLP)

Natural Language Processing (NLP) is a subset of Artificial intelligence (AI) which is used in game playing, speech recognition, understanding natural language processing, and computer vision. Salem (2000) notes that the research in NLP includes corpus-based methods, discourse methods, formal models, machine translation, natural language generation, and spoken-language understanding.

According to Dikli, NLP is a complex task to understand because it involves several levels of processing and subtasks and has four categories of language tasks including speech recognition, syntactic analysis, discourse analysis, information extraction, and machine translation.

AES AND FORMATIVE AND SUMMATIVE EVALUATION

In paper and pencil and computer delivery modes, assessments can be formative or summative. Formative evaluation can occur multiple times during the learning process. Its purpose is to provide understandable feedback to the learner and to guide instruction by the teacher. To be most effective, the feedback needs to be immediate, detailed and specific. It should be available in individual and group formats for the teacher, and it should suggest further directions for learning. On the other hand, summative evaluation is a one-time assessment for evaluating the skills and knowledge acquired at a specific reporting point on attained achievement at the student, school, district, and/or province/state level in order inform a decision about the individual or cohort. In many cases, the individual learner never receives feedback about performance or if so, it is in generalized terms, for example the student minimally met, successfully met or exceeded expectations, or the student successfully met or did not meet the assessment standard.

---

67 Haswell, Richard (2006). Automatons and Automated Scoring, p. 62.
68 Dikli, Semire. (2006). p. 7.

Summative and formative assessments place different demands on AES systems. The goal of summative assessment is to provide an accurate student score on one or more responses to writing prompts for the purpose of reporting. In this case, student responses can be selected for expert scoring which would then be used to train the AES system. The remaining batch of student responses could be submitted to the AES system for scoring. None of these stages requires an interactive response from the system; all can be done in scheduled batch mode. While there is considerable evidence that AES systems agree with human scorers on overall categorization of a student's responses, there is little evidence that any of the AES systems can demonstrate such agreement on individual attributes or features of writing. These attributes may include spelling, grammar, punctuation, tone, theme development and conclusion among many others. Even if there is a requirement for the AES system to provide information on attributes of writing, in summative systems, this is not likely to be required at the student level. There may be a greater tolerance for errors if reporting on attributes is limited to the system level.

Formative assessment places several different constraints on the AES system as the goal is immediate or near-immediate response to inform a student not only about the level of writing that was submitted but also about strengths and weaknesses of various attributes. The immediacy of the response implies that the service should be Internet based so that the student essay can be submitted directly to the AES system. A further implication is that the AES system needs to be already trained on the prompt to which the student responded. In turn, this means that educators should determine whether the available prompts match the types used in the classroom. To be of maximal use to the student, more information than a score on the essay must be provided: strengths and weaknesses on important attributes of writing are required so that the student can adjust or practice for future essays. Not all AES systems are capable of reporting on a set of attributes. Even systems that can report on individual attributes for a summative application have the additional burden of demonstrating that the confidence level for reporting on attributes at the school or jurisdictional level is adequate for the individual student.

The need for AES systems to demonstrate validity against a number of criteria is even more pronounced if it is to be used for formative purposes. A teacher, in reacting to a student's writing, clearly knows if the student has a grasp of a given attribute or if the student's response does not provide sufficient evidence for a comment about the attribute. Moreover, the teacher knows if the attribute is of importance for that prompt in the educational context of the classroom. At this point, few AES systems have demonstrated attribute importance and reliability in reporting at the student level.

MODELS FOR SUMMATIVE WRITING ASSESSMENT

Most AES programs were originally developed for summative evaluation purposes. As has been stated earlier, these purposes included differentiating among individuals for admission to selective programs, professional advancement, licensing, and certification. Large-scale, high-stakes assessments often also use summative AES evaluation programs. In the next section, brief overviews of IntelliMetric™, e-rater®, Project Essay Grader™ (PEG), Intelligent Essay Assessor™ (IEA), Bayesian Essay Test Scoring System (BETSY) are provided. There are other commercial and private instruments developed or in development besides the ones described but these are the most prevalent ones used in North America at this time.

Bayesian Essay Test Scoring sYstem™ (BETSY)

BETSY, developed by Lawrence Rudner with funds from the U.S. Department of Education and the Maryland State Department of Education, classifies text based on trained material (http:www/edres.org/betsy/help.htm). It was designed for automated essay scoring but can be applied to any text classification task. Based on the naïve Bayes text classification literature, its free software is downloadable and available for use.[69]

Project Essay Grader™ (PEG)

Project Essay Grader™ (PEG) was first developed by Ellis Page in 1966 with initial funding from the College Board. Valenti et al (2003) describe that PEG is one of the earliest and longest-lived implementations of automated essay grading. This program primarily relies on style analysis of surface linguistic features of a block of text, or in other words, it predominantly grades on the basis of surface linguistic features, taking no account of content. It is sometimes criticized for this focus on surface structures and ignoring of the semantic aspect of essays.

Project Essay Grader™ uses regression coefficients calculated from training essays marked by human raters to predict the intrinsic quality of the essays to be scored. Valenti et al. explain the PEG model as being based on the concept of "proxes" or computer approximations or measures of "trins" that are intrinsic variables of interest in the essay. Trins are what human raters would search for in an essay but which the computer does not directly measure. Examples of the relationship between proxes and trins would be: proxes—essay length indicates the trin—fluency; proxes—counts of prepositions, relative pronouns and other parts of speech indicates trin—sentence structure; proxes—variation in word length indicates trin—diction.

From its early days Page stated that PEG could predict scores that were comparable or better than individual human raters. In 1993 it was modified in several significant ways when the program acquired several parsers and various dictionaries, and also incorporated special collections and classification schemes. By 2001 a study scoring essays using the World Wide Web found that the speed of marking was about three essays every second and the cycle time for the submission of the essay to finished report was about two minutes.[70]

PEG requires 100 to 400 "previously manually marked essays for proxes, in order to calculate the regression coefficients, which in turn enables the marking of new essays."[71]

e-rater® (Educational Testing Service ETS)

Electronic essay rater or e-rater®, was first developed by Burstein and Kaplan at ETS (Educational Testing Service) in the 1990s and was used beginning in 1999 to score the Graduate Management

69 Rudner, Lawrence and Gagne, Phill. (2001). An Overview of Three Approaches to Scoring Written Essays by Computer. *Practical Assessment, Research & Evaluation.* 7 (26), p. 5.

70 Page, Ellis Batten (2003). Project Essay Grade: PEG. p. 50.

71 Valenti, Salvatore, Nitko, Anthony J., and Cucchiarelli, Alessandro. (2003). p. 321.

Admissions Test Analytical Writing Assessment (GMAT AWA). Most of the literature is written about the first version of this AES engine; however, according to Attali and Burstein (2006) in the second version "one of the most important characteristics of e-rater® V.2 is that it uses a small set of meaningful and intuitive features. This distinguishing quality of e-rater® allows further enhancements that together contribute to a more valid system."[72]

Rudner and Gagne explain that the first e-rater® utilizes "sophisticated hybrid feature technology that uses syntactic variety, discourse structure (like PEG) and content analysis (like IEA). To measure syntactic variety, e-rater® counts the number of complement, subordinate, infinitive, and relative clause and occurrences of modal verbs (would, could) to calculate ratios of the syntactic features per sentence and per essay. For structure analysis, e-rater® uses 60 different features, similar to PEG's proxes." [73] These 60 different features have been reduced and the feature set for e-rater® V.2 now "includes measures of grammar, usage, mechanics, style, organization, development, lexical complexity, and prompt-specific vocabulary usage."[74]

Both versions of e-rater® use natural language processing, which is a subset of AI or artificial intelligence. It evaluates the quality of an essay by identifying linguistic features in the text. e-rater® employs a corpus-based approach to model building which usually requires researchers to use copy-edited text sources like newspapers. However, as Dikli explains, unedited text corpora representing the particular genre of first-draft student essays is required by e-rater®'s feature analysis and model building.

Intelligent Essay Assessor™ (Pearson Knowledge Analysis Technologies)

The Intelligent Essay Assessor (IEA) uses a text-analysis method called Latent Semantic Analysis (LSA) created by psychologist Thomas Landauer with the assistance of Peter Foltz and Darrell Laham. IEA is a product of Pearson Knowledge Analysis Technologies. The program's main focus is more on the content-related features (quality of content) rather than formulated ones; however, IEA does include scoring and feedback on grammar, style and mechanics as well as validation and plagiarism detection. According to Landauer et al. (2000) when comparing AES systems they note:

> Other systems work primarily by finding essay features they can count and correlate with ratings human graders assigned. They determine a formula for choosing and combining the variables that produces the best results on the training data. They then apply this formula to every to-be-scored essay. What principally distinguishes IEA is its LSA-based direct use of evaluations by human experts of essays that are very similar in semantic content. This method, called *vicarious human scoring,* lets the implicit criteria for each individual essay differ.[75]

---

72 Attali, Y. and Burstein, J. (2006). Automated essay scoring with E-rater® V.2. *Journal of Technology, Learning, and Assessment* 4 (3), p. 7.

73 Rudner, Lawrence and Gagne, Phill. (2001). p. 3.

74 Attali, Y. and Burstein, J. (2006). p. 7.

75 Landauer, Thomas K., Laham, Darrell, and Foltz, Peter W. (2000). The Debate on Automated Essay Grading: The Intelligent Essay Assessor. *IEEE.* p. 28.

Landauer et al (2003) claim that IEA can successfully analyze not only content-based essays but also creative narratives. The system needs to train on a set of domain-representative texts in order to measure the overall quality of an essay. Landauer et al. (2003) claim that the system needs smaller numbers of pre-scored essays to train. Unlike other AES systems requiring 300 to 500 training essays per prompt, IEA only requires one hundred pre-scored essays.

IntelliMetric™ (Vantage Learning)

IntelliMetric™ was developed by Vantage Learning as the first essay-scoring tool that was based on artificial intelligence (AI), specifically NLP (Elliott, 2003; Shermis & Barrera, 2002; Shermis, Raymat, & Barrera, 2003). IntelliMetric™ was developed by Vantage Learning and used by the College Board for placement purposes (Myers, 2003).

Starting January 2006 ACT™, Inc. is responsible for GMAT test development and scoring. IntelliMetric™ by Vantage Learning was part of the initial proposal to the Graduate Management Admission Council®. It is used to rate the Analytical Writing Assessment (AWA) section of the GMAT. The two 30-minute writing tasks, Analysis of an Issue and Analysis of an Argument, are designed as a direct measure of the candidate's ability to think critically and communicate ideas. Examinees use rudimentary word processing functions for the final product, but also have available scratch paper or erasable note boards. Two human raters using detailed scoring rubrics initially score the prompts. In case of disagreement of more than one score on a six-point scale, a third human rater is used. Based on a sufficient number of prompts an automated essay scoring model is developed and evaluated and for the remaining papers IntelliMetric™ and one human rater grade the majority of essays.

Dikli summarizes the five key principles of the IntelliMetric™ system as being *neurosynthetic* or modeled on the human brain, a learning engine, systemic and based on a complex system of information processing, inductive, multidimensional and non-linear.[76]

Rudner et al.[77] state that Vantage Learning's corporate strategy appears to be to protect the IntelliMetric™ system by treating details of the technology as proprietary trade secrets. Though many patents and specifics are not disclosed, Rudner et al. believe a paper by Elliott and Mikulus (2004) provides insight into the IntelliMetric™ system. IntelliMetric™ combines scores from focus and unity (coherence), organization, development and elaboration, sentence structure, mechanics and conventions to attain a final score.

IntelliMetric™ evaluates essay responses in multiple languages including English, Spanish, Hebrew, and Bahasa, and evaluation of text can occur in Dutch, French, Portuguese, German, Italian, Arabic, and Japanese.[78]

---

76 Dikli, Semire. (2006). p. 17.

77 Rudner, L. M., Garcia, V., and Welch, C. (2006). p. 6.

78 Elliott, Scott (2003). Intellimetric™: *From Here to Validity in Automated Essay Scoring: A Cross-Disciplinary Perspective*. Edited by Shermis, Mark D. and Burstein, Jill. Lawrence Erlbaum Associates: Mahwah, New Jersey, p. 72.

Elliott describes the system as being able to be applied in "Instructional" (formative evaluation - MY Access!®) or "Standardized Assessment" (summative evaluation - IntelliMetric™) modes. The Standardized Assessment mode typically provides a holistic score and, if appropriate, feedback on various rhetorical and analytical dimensions of the writing sample essay.[79]

MODELS FOR FORMATIVE WRITING ASSESSMENT

While most AES programs were originally developed for summative evaluation purposes, formative evaluation programs can be very useful in writing instruction and as Shermis and Burstein (2003) note the credibility of AES systems will increase when their use includes more formative assessment. Students can submit multiple revisions for assessment, and based upon the assessment results, revise their work so that it becomes more proficient, ideally both in content and style. AES programs can provide feedback directly to the student without teacher intervention (non-mediated), the teacher can mediate the computerized feedback or in some programs the teacher can augment the feedback provided by the computer program. In this section, brief overviews of MY Access!® and Criterion℠ are provided. As in the summative evaluation application section, there are other instruments available or in development besides the two described.

Criterion℠ (Educational Testing Service ETS)

Criterion℠ is a web-based formative writing assessment tool which uses two ETS technologies, e-rater® and Critique. Burstein (2003) states that the advisory or feedback component of Criterion℠ supplements but does not determine the writing score.[80] The advisory component contains "feedback to indicate the following qualities of an essay response: (a) the text is too brief to be a complete essay (suggesting that the student write more), (b) the essay text does not resemble other essays written about the topic (implying that perhaps the essay is off-topic), and (c) the essay response is overly repetitive (suggesting that the student use more synonyms)."[81]

Programs included in Critique provide holistic scoring, and individualized diagnostic feedback. They detect errors in grammar, usage, and mechanics and recognize discourse elements and elements of undesirable style in an essay. ETS claims Criterion℠ can assess a number of writing genres including persuasive, descriptive, narrative, expository, cause and effect, comparison and contrast, problem and solution, argumentative, issue, response to literature, workplace writing, and writing for assessment. e-rater® models exists for grades 4 through 12 as well as for undergraduates. Prompts are from national standards, English Proficiency Test, PRAXIS, TOEFL, GMAT, and GRE. In addition to these prompts, teachers can create and assign their own topics. Feedback of every dimension of writing except holistic scoring can be reported for teacher-created topics. (ETS, n.d.).

Included in Criterion℠ are student tools such as a writer's handbook and an electronic portfolio which allows students to archive drafts and see feedback definitions, examples of correct and incorrect

79 Ibid., p. 72.
80 Burstein, J. (2003).
81 Ibid., p. 119.

use, and error explanations. Teacher options include the ability for teachers to add their own feedback, set start/finish dates for students, and control student access to features such as spell check, diagnostic feedback or holistic marking.

ETS claims Criterion℠ can be used for instruction, remediation, placement, and benchmark and exit testing.

MY Access!® (Vantage Learning)

Vantage Learning markets MY Access!®, which is a web-based formative writing assessment tool using IntelliMetric's℠ automated essay scoring engine. Students are offered a writing environment that provides immediate scoring and diagnostic feedback with or without teacher intervention. This feedback helps students revise their essays, and Vantage Learning claims this motivates them to improve their writing proficiency. [82]

According to Vantage Learning, MY Access!® provides diagnostic information about students, ensures greater grading consistency and accuracy among teachers and schools, and allows teachers more time for data-driven curriculum planning, and decision making to conduct differentiated instruction.

Like IntelliMetric™, MY Access!® is multilingual which is very helpful especially with English as second language students in grades 4–12 and at the college level. There are two options: (1) the student can write on a topic in English, Spanish, or Chinese and receive feedback in the same language; or (2) the student can write an essay in English and can receive feedback in either English or their own language. Some features include: multilevel feedback—developing, proficient, and advanced; multilingual dictionary thesaurus; and translator functions. Vantage Learning plans to expand the language options in the future.

MY Access!® has over 200 operational and pilot prompts based on reading texts as well as literature at different educational levels. While teachers could provide their own prompts, the system will be unable to score the student essays until it first is trained on about 300 other human scored essays. It claims to be able to provide feedback on different genres of writing, such as informative, narrative, literary, and persuasive essays.

Included in MY Access!® are student tools such as *writing dashboard* and *my portfolio* which allows students to archive work and see their weekly progress. Teacher options include the ability for teachers to add their own feedback, generate up to 10 different student progress reports, and produce multilingual parent letters.

Other Models

As stated earlier, there are other models for writing assessment. Valenti et al. provide descriptions of Conceptual Rater (C-Rater), Automark, and SEAR, Paperless School free text Marking Engine[83]. In

82 Vantage Learning (2006, August). MY Access®. Retrieved from http://www.vantagelearning.com/myaccess/

83 Valenti, Salvatore, Nitko, Anthony J., and Cucchiarelli, Alessandro. (2003). pp. 324-326.

Australia, Williams and Dreher describe a program being developed by Curtin University researchers, MarkIT[84], which they define as a simple but comprehensive form of feedback to essay authors which enables the authors to see where essay content is inadequate in terms of discussion of the essay topic. In the United States several programs are in different stages of development and testing, e.g., SAGrader, WriteToLearn[85], and ETIPS[86].

COMPARISON OF AUTOMATED ESSAY SCORING SYSTEMS

Practitioners and decision makers who are evaluating the use of AES may find Table 3.1 Comparison of AES Systems (Dikli, 2006) useful. Attali and Burstein (2006) believe that AES systems do not directly evaluate the intrinsic qualities of an essay as human raters do, but instead AES systems predict the score of an essay by using correlations of the intrinsic qualities. Each of the AES systems in the chart employs various techniques: e-rater® and IntelliMetric™ use NLP techniques; IEA™ uses LSA; PEG™ utilizes proxy measures (proxes) and BETSY™ uses Bayesian methods. Criterion℠ is the instructional application based on e-rater® and MY Access!® is based on IntelliMetric™. The other programs do not yet have a formative assessment program. The numbers of essays needed to train each system is noted in the table.[87]

Table 3.1 Comparison of AES Systems

AES System	Developer	Technique	Main Focus	Instructional Application	Number of Essays Required for Training
PEG™	Page (1966)	*Statistical*	Style	N/A	100–400
IEA™	Landauer, Foltz, & Laham - 1997	*LSA*	Content	N/A	100–300
E-rater®	ETS development team (Burstein, et al., 1998)	*NLP*	Style and content	Criterion℠	4
IntelliMetric™	Vantage Learning (Elliott, et al., 1998)	*NLP*	Style and content	MY Access!®	300
BETSY™	Rudner (2002)	*Bayesian text classification*	Style and content	N/A	1000

84 Williams, R. and Dreher, H. (2005). Formative Assessment Visual Feedback in Computer Graded Essays. *The Journal of Issues in Informing Science and Information Technology* 2.

85 Pearson Knowledge Technologies, WriteToLearn. Retrieved from http://www.pearsonkt.com/prodWTL.shmtl

86 Riedel, E., Dexter, S., Scharber, C., & Doering, A. (2005). Experimental Evidence on the Effectiveness of Automated Essay Scoring in Teacher Education Cases. *86th Annual Meeting of the American Educational Research Association*, p. 13.

87 Dikli, Semire. (2006).

Other differences between the systems not noted on Table 3.1 are that currently IntelliMetric™ is the AES model with the most diverse multilingual capacity and that BETSY is the only free and readily downloadable software.

From the description of the various AES engines above it is apparent that they are driven by different assumptions which result in different implications for formative or summative evaluation usage. However, not enough information about the assumptions inherent in the systems is readily available for comparative analysis.

DIFFICULTIES WITH DIRECT SCIENTIFIC COMPARISONS

To date there seems to be a dearth of independent comparative research on the effectiveness of the different AES engines for specific purposes, and for use with specific populations. It is very difficult to ensure that a transparent scientifically-based comparison can be made among the major AES engines due to the proprietary nature of many of the scoring engines, their differing focuses, and their differing uses. Some AES engines concentrate on style, others on content, and still others purport to score both style and content. While it would appear that one basis of comparison might be the degree of agreement of specific AES engines with human raters, this also needs to be scrutinized as different prompts, expertise of raters, and other factors can cause different levels of rater agreement. Recommendations as to further research and examples of some of the issues to be addressed in independent comparative research if AES is to be implemented in Canadian K–12 educational settings (e.g., classrooms, online learning, large-scale assessments) are indicated in the final chapter of this paper.

4. Reliability and Validity

AES is an immature, emerging technology with much potential but several threats to validity that need to be studied and dismissed.[88]

This chapter presents information related to the reliability and validity of AES. Dikli cites a number of studies that have been conducted to assess the accuracy and reliability of the AES systems with respect to writing assessment. The results of several AES studies reported high agreement rates between AES systems and human raters.[89]

Keith (2003)[90] discusses the general types of evidence that AERA et al. (1999) state are relevant for validity including evidence based on test content (content validity), internal structure (internal validity), relations to other variables (external validity), and the consequences of testing. He then comments that "although AES systems have just scratched the surface of demonstrating such evidence, the standards and traditional definitions of validity provide a categorization for validity evidence that has been gathered and a blueprint for further studies."[91]

For content evidence Keith notes that AES systems represent scoring systems rather than tests and as the content of essays is independent of the method of scoring, that content validity evidence is not particularly relevant for AES systems.

The central question for AES systems and the nexus of questions from skeptics of AES according to Keith is whether AES scores reflect writing skill or some other characteristic. Some of these characteristics might be general cognitive ability, content vocabulary knowledge or simply the ability to produce a large amount of text in a limited time. Alternatively, he suggests the results could reflect simple fantasy, with the scores having no real meaning. This, of course, is also a central question for human marking. Keith continues that "most AES programs have implicitly or explicitly assumed that human raters are indeed able to score prose for general writing skill or content-specific writing skill with some degree of validity."[92] If one assumes the scores from human raters lead to valid inferences

88 Haladyna, T. M. & Olsen, R. M. (2006). *Threats to Validity in Large-Scale Writing Performance Tests: What Are These Threats and What Should Be Done About It*. p. 17.

89 Dikli, Semire. (2006). An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning, and Assessment* 5 (1), p. 4.

90 Keith, Timothy Z. (2003). *Validity and Automated Essay Scoring Systems in Automated Essay Scoring: A Cross-Disciplinary Perspective.* Edited by Shermis, Mark D. and Burstein, Jill. Lawrence Erlbaum Associates: Mahwah, New Jersey, p. 147.

91 Ibid., p. 147.

92 Ibid., p. 148.

of writing skill, correlations of AES program scores with those of human raters demonstrate that the AES system also measures writing and that "correlations may also legitimately be considered evidence of reliability and criterion-related validity."

The scores of human raters are based on rubrics that articulate the characteristics of writing that are valued for a given writing task. Using correlations of AES scores with the scores of human raters as evidence of validity is problematic in at least two ways. Human scores are based directly on the application of rubrics describing each scoring category together with sets of essays that have been designated by experts as exemplifying each of the categories (exemplars). Challenges to the scoring of a particular essay are resolved by reference back to the rubrics. AES scoring engines are trained on a large set of essays scored by experts. The engine is thus one step removed from the rubrics which define the required qualities of writing. This may be less of an issue with engines that have regression variables based on the characteristics reflected in the rubric.

The second issue is that each essay topic requires a new set of papers for the training of the AES engine. This implies that the selected regression variables and weights are specific to the topic which is akin to having a slightly different rubric applied to each essay topic. It may also be that the selected regression variables and weights reflect differences between the panels of experts that select the training papers for different administrations rather than differences in the way that students respond to the different topics. This is a challenge to the validity of scoring.

Keith states that exploratory and confirmatory factor analysis of AES scores with other measures of writing may be another method of establishing whether AES measures the same or divergent constructs. As AES programs increasingly score components of writing such as content and mechanics, factor analysis of the component scores can help demonstrate the constructs have been measured. In addition, he postulates that we would expect valid AES scores to correlate with standard achievement tests having a higher correlation with reading and writing scores and a lower correlation with mathematics or science achievement results.

Criterion related validity has many potential areas with which AES scores could be correlated, e.g., achievement test scores, Graduate Record Exam results, writing performance in classrooms and classroom examinations.

Keith brings forward a number of issues to be considered when evaluating or conducting AES research. The first of these is the issue of calibration or validation. Most AES engines are trained on a sample of essays that have been scored by human raters. The number of essays required for this process varies from program to program. Elliott[93] describes that this training uses multiple regressions and chooses a set of predictor variables and optimal regression weights for predicting the ratings for a human judge or the averaged ratings of more than one judge, which are then used to score essays that have not been previously scored by humans. Keith cites Burstein, Kukich, Wolff, Lu and Chodorow (1998) stating that a common variation in the process Elliott outlines is to have multiple judges

93 Elliott, Scott (2003). Intellimetric™: From Here to Validity. pp. 71-72.

involved in training but only a single human rater and the AES engine used in subsequent scoring. Correlations between human raters and AES engines will vary depending if the calibration sample is used to determine correlations or if a separate cross validation sample is employed. Keith cautions against inflated estimates of validity if there is not a separate cross validation sample.

Haladyna and Olsen (2006)[94] presented to the CCSSO Large-Scale Assessment Conference a paper that outlines threats to validity for large-scale writing performance tests. The concepts discussed in their paper are summarized in this literature review because these threats to validity for large-scale writing performance tests are key factors to consider for any type of writing assessment whether it be administered in a paper and pencil or a computer-assisted delivery mode. After presenting background ideas to provide a context for a discussion of validity as it bears on measuring writing ability, they identify, describe, and document threats to validity and present some recommendations about what test developers and sponsors of writing performance testing programs should do to lessen or eliminate each threat. Some of the potential threats to validity are analyzed in Table 4.1.

Table 4.1 Potential Threats to Validity in Writing Assessments [95]

Validity Evidence	Salient Questions
Content-Related	What is writing ability? How does computer ability or handwriting legibility affect the writing construct? How do the students' emotions interplay in measuring writing ability? Which writing tests have greatest fidelity with this definition? Is a single prompt a sufficient sample for a writing performance test? What is the structure and validity of analytic traits?
Reliability	What is the reliability estimate for a set of writing test scores? What is the standard error of measurement around the cut score? How consistent are raters? What is the reliability of any classification?
Prompt Quality	Are prompts validated? How representative are prompts of the writing construct? Are prompts field tested? Are the rubrics adequately developed? Do prompts have differential item functioning?
Test Design	Does the test follow its test specifications? How does choice of prompts operate in this design? Are students given an opportunity to maximize their performance? If writing can be improved by using the computer, should students have this opportunity? Does the test design permit writing in a prescribed manner consistent with sound instructional practices that are recommended in the sponsoring unit (state or school district)?
Administration	Are writing tests administered in a standardized way?
Scoring	Are raters adequately trained? Have scoring errors occurred? Is there adequate quality control? What about rater effects, such as severity/leniency, central tendency idiosyncrasy, restriction of range, halo, and logical errors?
Scaling for Comparability	How can we place all prompts on the same test score scale?
Standard Setting	If different standard-setting methods produce different recommendations for cut scores, is there a "true" cut score?

94 Haladyna, Thomas M. and Olsen, Robert M. (2006).
95 Ibid., p. 10.

Haladyna and Olsen make several important points. The first being that writing test scores have many different purposes: to inform teachers to be able to plan instruction; to inform students; to meet the need for accountability; and to guide policy analysis and funding. They state that each and every use of a writing test score requires separate validation. These validations should be done by an experienced, knowledgeable independent evaluator. The second point is that though performance testing will likely provide greater challenges to validation than SR (selected response) testing, this is not an argument in support of SR testing and against performance testing but simply recognition that threats to validity are probably greater with performance testing. [96]

Haladyna and Olsen state that scoring student essays is troublesome because of potential threats to validity. Some of the threats that they document include:

- **Rater effects.** Biases include rater severity, central tendency, restriction of range, halo and idiosyncrasy. They refer readers to Myford and Wolfe (2003). Rater severity can account for up to 13% of the variance in ratings.[97]
- **Training to Reduce Rater Effects**. Review of research on the effectiveness of training shows that it has not reduced rater effects sufficiently (Haladyna, 2004a) . . . no training model eliminates rater effects and perforce. The problem remains to be solved.
- **Score Resolution to Eliminate Rater Inconsistency**. While score resolution is frequently used in high-stakes, large-scale performance testing to increase fairness when raters do not agree after evaluating a student paper, the benefit of score resolution is to increase rater consistency, which improves reliability not validity. Haladyna and Olsen state "resolution only deals with random error; it does not eliminate systematic error that is associated with CIV. Two raters may be in agreement but extremely harsh or extremely lenient. In such instances, both raters are wrong in their judgment, and score resolution methods do not identify and correct this problem."
- **Detection of Faked Essays**. Automated Essay Scoring (AES) may not detect faked essays even though Powers, Burstein, Chodorow, Fowles, & Kukich (2002) reported a study on the ability of one application of AES to detect different types of faked essays. However, many teachers are still suspicious that nonsense essays could earn average or higher scores.
- **Wordiness**. Powers (2005) reviewed the literature on the relationship between essay length and overall score. An increase in essay length can increase one's test score; therefore, it is essential that any validation of a writing performance test guarantees that essay length does not overly influence the overall score.
- **Scoring Errors.** Haladyna and Olsen state most of the documented cases of scoring errors by scoring companies involve SR (selected response) tests but they caution that performance tests such as writing tests are not immune to scoring errors. "With respect to writing performance tests, if any student score does not correspond with other achievement data about that student's writing, then the validity of that score is in doubt. Since by design, almost all state writing tests are one-trial performances, no other information is obtained to corroborate a student's writing performance or to question a score's validity. Without such safeguards,

96 Haladyna, Thomas M. and Olsen, Robert M. (2006), p. 8.
97 Ibid., p. 17.

students have no protection against scoring errors or other irregularities. Schools and school districts run similar risks if they are accountable for student performance and scoring errors affect a large unit of analysis, such as a school or school district."[98]

- **Scaling for Comparability.** Uniform scales that span grade levels and years or measure adequate yearly progress and growth through grade levels are needed. Horizontal scaling is required if multiple test forms are used in the same year, and vertical scaling is required if growth over time through grade levels is to be measured.
- **Standard Setting**. While the science of standard setting is well established for SR tests, it is not as well developed for performance tests (Cizek, 2006).

Haladyna and Olsen conclude their paper by stating that "we have shown far too many threats to validity in a writing performance test to warrant its use in high-stakes, large-scale testing . . . until all threats have been identified, studied, and resolved, high-stakes use of writing test scores is unwise. Toward that end, some recommendations seem justified."[99] The recommendations they advise include:

- **Independent Evaluations (Audits)** The virtue of an independent evaluation of a testing program is that a test sponsor/developer gets the benefit of an independent opinion about validity.
- **Validity Studies.** Validity studies can address principles or procedures of a specific testing program or generically address problems that affect virtually all performance testing programs, such the problems posed with rater effects.
- **Technical Reports**. The technical report should organize evidence that links to AERA *Standards* (AERA, et al., 1999). Toward that end, they present an outline for an ideal technical report with leading questions driving the evidence that should be assembled to support validity and call attention to each threat to validity.

The authors believe that not enough evidence is presented in public documents to support the use of writing test scores in high-stakes, large-scale testing programs in many of the intended ways.

RATER AGREEMENT

Yang (2002) notes there are three methods for system validation: (1) single essay agreement results with human scores, (2) correlations between scores on different prompts, and (3) descriptions of the scoring process and how it contributes to the validity of the system.[100] Many studies have been conducted with different AES engines using the first validation method—single essay agreement results with human scores. AES engine vendors and others have been active in trying to provide this type of validity. Valenti et al. (2003) in their overview of some of the major AES engines provide a brief performance statement for each, parts of which are cited below:

- PEG: Page's latest experiments achieved results receiving a multi-regression correlation as

98 Ibid., p. 19.

99 Haladyna, Thomas M. and Olsen, Robert M. (2006), p. 20.

100 Yang, Yongwei, Buckendahl, Chad W., Juszkiewicz, Piotr J., and Bhola, Dennison S. (2005). Evaluating Computer Automated Scoring: Issues, Methods, and an Empirical Illustration. *Association of Test Publishers Journal*. July 2005.

high as 0.87 with human graders.

- IEA: A test conducted on GMAT essays using the IEA system resulted in percentages for an adjacent agreement with human graders between 85% to 91%.
- E-rater: over 750, 000 GM ET essays have been scored, with agreement rates between human expert and system consistently above 97%. By comparing human and e-rater grades across 15 test questions, the empirical results range from 87% to 94%.
- BETSY: An accuracy of over 80% was achieved with the described data set.[101]

Dikli also concurs that correlations and agreement rates between the AES engines described and expert human raters are high.

While these appear at first glance to be impressive and may be, there are other issues such as exact and adjacent agreement and expert versus standard human scoring to consider. Dikli discusses these issues and states that first it is critical to understand the difference between exact agreement and adjacent agreement. "Exact agreement requires two or more raters to assign the same exact score on an essay (e.g., two raters assign 5 on a 1–6 scoring scale). On the other hand, adjacent agreement requires two or more raters to assign a score within one scale point of each other (e.g. one rater assigns 5 and another rater assigns 6 respectively on a 1–6 point scoring scale). It is clear that exact agreement is harder to achieve and that adjacent agreement results in higher agreement rates."

In Table 4.2 Dikli combines the concepts of exact and adjacent agreement and expert versus standard human raters. This table compares agreement rates for expert scoring, standard human scoring and IntelliMetric™ scoring when assessing grade eight students in a statewide testing program. The data illustrates that the adjacent agreement rates between humans, and IntelliMetric™ and humans can be higher than the exact agreement rates. The methodology was that two expert raters scored each essay followed by two standard human scorers who independently scored the writing, and then IntelliMetric™ scored each essay.

Table 4.2 Comparison of Expert Scoring, Human Scoring, and IntelliMetric™ Scoring[102]

	Human 1 to Human 2	Human 1 to IntelliMetric	Human 2 to Intellimetric	Human 1 to Experts	Human 2 to Experts	IntelliMetric to Experts
Exact	.52	.53	.56	.58	.54	.73
Adjacent	.94	.96	.95	.96	.97	.99
Discrepant	.06	.04	.05	.04	.03	.01

Dikli states for this study "expert" referred to an individual who had a degree in English as well as at least five years of experience in analyzing writing in large-scale writing assessment programs and "traditional human scorer" was defined as an individual who usually attended a one-day training session

101 Valenti, Salvatore, Nitko, Anthony J., and Cucchiarelli, Alessandro. (03). An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education* 2 (Information Technology for Assessing Student Learning Special Issue), pp. 321-326.

102 Dikli, Semire. (2006). p. 27.

on writing assessment in a large, statewide scoring session in writing (Vantage Learning, 2002).

In many scoring situations there is controversy over who is qualified to be a human scorer. While some jurisdictions and/or companies may require "experts" and also provide training, some are being accused of less rigorous guidelines. The NCTE Task Force on SAT and ACT Writing Tests claims that the "gap between the SAT essay score and college course may be further widened by the way the essay is scored, because Pearson Educational Measurement, which has been hired by the College Board to rate the essays, does not require their raters to have taught in college; the College Board reports that 58% of its current reader pool teaches at the postsecondary level and 42% at the secondary level."[103]

In Texas it is claimed "scorers spend an average of 2 1/2 minutes deciding whether a student's effort passes or fails. And that has a growing number of critics, including testing experts and legislators in other states, wondering if the part-time workers, who make as a little as $11 an hour, are shortchanging quality for speed . . . In Texas, the TAKS scorers are hired by a contractor who requires them to sign confidentiality agreements that prohibit them from talking about how the system operates. They must have a bachelor's degree in any subject, but they are not required to have experience in education."[104] According to this article the same concerns have arisen in Florida, particularly in short-answer questions in subject areas where the rater may not have expertise. Of the 1.7 million essays written by Texas students last year scored by 1,950 workers with each scorer reading 150 tests per day, there were only 1,044 challenges lodged by parents and districts for low scores in 2006 and only 65 scores changed as a result of challenges in 2006.[105]

In this chapter information related to the validity and reliability of AES systems was discussed. While there are many cautions cited related to validity, it appears that the AES engines profiled have high single essay agreement results with human scores. Where research has been done, it appears that AES engines can have greater exact agreement with expert human raters than they do with standard human raters. Some of the major difficulties in trying to make comparisons between different AES programs in regards to validity rates are: differences in the type of human rater used; the use of AES scoring tools, and the focus of the programs, for example; some programs focus on content and style – others on one or the other. To have a true comparative study, the same students would need to respond to the same prompt under the same parameters but be scored by each of the AES engines. This type of research does not appear to have occurred to date.

---

103 NCTE Task Force on SAT and ACT Writing Tests (2005, April). *The Impact of the SAT and ACT Timed Writing Tests: Report From the NCTE Task Force on SAT and ACT Writing Tests.* pp. 3-4. Retrieved from http://www.ncte.org/library/files/About_NCTE/Press_Center/SAT/SAT-ACT-tf-report.pdf

104 Berard, Yamil and Cromer Brock, Katherine (2006).

105 Ibid., p. 1.

5. AES: Implications for K-12 Pedagogy

The use of computer technologies has the potential to make fundamental changes in how we teach, which mental processes, skills and understandings we measure, and how we make decisions about student learning.[106]

Taylor states that radical change is occurring in the philosophy, intent and practice of assessment. Traditionally assessment was a post-instruction measure of learning but this has changed to include processes tightly linked with instruction in order to increase student achievement. Taylor notes that it is important to distinguish between assessment of learning and assessment for learning. Assessment of learning (summative) is when the assessment tool provides information to help make final judgements of competency or to make performance comparisons among jurisdictions. Assessment for learning (formative) identifies areas of strength or weakness for use as a tool in gaining direction for instruction or remediation.[107]

As has been stated earlier, the use of AES is growing exponentially in recent years. While it has been primarily used as an assessment of learning or a summative evaluation tool, it has recently been adapted to also be used as an assessment for learning or a formative evaluation tool. This follows the general trend in all assessment as Taylor states that formative evaluation is the fastest growing component of the testing industry in recent years.[108]

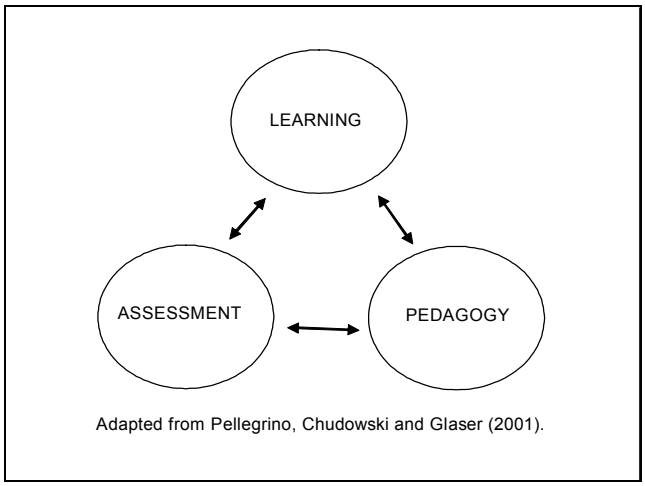
While the use of AES is growing exponentially, research about the pedagogical issues related to it has yet to keep pace. Most research to date has been related to justification of the AES programs and researchers are only now starting to look at questions surrounding its use in the classroom. This is partially because until the last few years AES was used mostly in large-scale summative assessments at the college or university level.

This is unfortunate, for as Ridgway et al. (2004) discuss there is an intimate association between teaching, learning and assessment. This is illustrated in Figure 5.1, Learning, Assessment and Pedagogy created by Robitaille et al (1993). This diagram outlines the three components of the curriculum: the

106 Taylor, Alan R. (2006), p. 11.
107 Ibid., p. 17.
108 Ibid., pp. 18-19.

intended curriculum (set out in policy statements), the implemented curriculum (which can only be known by studying classroom practices) and the attained curriculum (which is what students can demonstrate at the end of a course of study). They state that the assessment system—tests and scoring guides—provides a far clearer definition of what is to be learned than does any verbal description and therefore, is a basis for curriculum planning at the classroom level. While they acknowledge teachers' values and competences also mediate policy and attainment, they deem the assessment system is the most potent driver of classroom practice.[109]

Figure 5.1 Learning, Assessment and Pedagogy[110]



THE TEACHING OF WRITING

The first principle cited by NCTE is that "everyone has the capacity to write, writing can be taught, and teachers can help students become better writers." The NCTE Beliefs Statement notes "what teachers do makes a difference in how much students are capable of achieving as writers." If AES is determined to be beneficial for student learning in both formative and summative assessment, then teachers will need to learn how to use this tool in a variety of situations. "Teachers of writing should be well-versed in composition theory and research, and they should know methods for turning that theory into practice." The NCTE statement continues "much as in doctoring, learning to teach well is a lifetime process, and lifetime professional development is the key to successful practice. Students deserve no less."[111]

As people learn to write by writing and NCTE suggests providing ample in-class and out-of-class opportunities for writing including writing for a variety of purposes and audiences, then AES needs to be developed further in order to allow it to assess writing that varies in form, structure, and production process according to its audience and purpose. Writing teachers discuss the tension between writing

109 Ibid., p. 5.
110 Ridgway, Jim, McCusker, Sean, and Pead, Daniel (2004), p. 5.
111 Writing Study Group of the NCTE Executive Committee (2004), p. 1.

as generating and shaping ideas and writing as demonstrating expected surface conventions. NCTE's stated policy is that conventions of writing are best taught in the context of writing. Simply completing workbook or online exercises is inadequate if students are not regularly producing meaningful texts themselves. If AES can be implemented in such a manner, especially in its formative assessment applications, students can learn the conventions of language in the context of writing, receiving feedback on many versions of the same writing piece and/or on many more samples of writing than a teacher currently can assess. Especially for those students who have difficulties with reading, AES may be able to be incorporated with reading assistance programs to strengthen the connection for those students between their reading and writing.

Teachers already incorporate increasingly rapid changes in technologies such that composing can involve a combination of modalities, such as print, still images, video, and sound as computers make it possible for these modalities to work together. As the NCTE states "from the use of basic word processing to support drafting, revision, and editing to the use of hypertext and the infusion of visual components in writing, the definition of what writing instruction includes must evolve to embrace new requirements." AES is an example of a new requirement.

Even though assessment of writing involves complex, informed, human judgment, some aspects of writing assessment can be performed well by AES already. Therefore, it would seem most beneficial to use AES to assess those components at which it excels and to use a combination of human and computer assessments to allow more opportunities for feedback to students. AES could be used to help writing teachers individualize the writing instruction even more than currently is possible due to time and resource restrictions. The skill of knowing how to deliver useful feedback, appropriate for the writer and the situation could be further refined by learning how and when to have students receive non-mediated formative assessment from AES, and when it is best for teachers to mediate in the process. This would involve the complex, informed human judgement about when the teacher is able to encourage autonomy in the student and when the student is not able to effectively use the information provided without teacher mediation.

EVALUATION OF WRITING

Formative Evaluation

As described in Chapter 3, AES engines have begun to be used for formative evaluation purposes where students can submit various drafts of their writing to be scored, and then choose to incorporate the feedback into the next draft of the writing. Criterion℠ from ETS and MY Access!® from Vantage Learning are major commercial programs already in use. WriteToLearn from Pearson Knowledge Technologies is a new commercial entry into this market for use from Grade 6 and above that uses the KAT (Knowledge Analysis Technology) engine. Other individuals or groups have also developed or are developing formative evaluation programs.

Formative evaluation can be provided in both non-mediated and mediated formats, for example, the student can receive the feedback and proceed to the next draft without teacher intervention or it can be in a mediated situation where the teacher controls a variety of factors. The teacher can decide on which method best meets the instructional needs of a specific individual, assignment or class.

Summative Evaluation

Taylor states that "among current applications used extensively in some jurisdictions are the following: final examinations to determine standing in a course, entry examinations to identify those who meet the criteria for entry into a subsequent level of education or training, credentialing examinations used to identify those who should be certified to practice in the professions and trades, and assessments to determine the effectiveness of programs."[112] Summative evaluations using AES engines as a co-rater in large scale assessments fulfill the types of purposes he lists. For example starting January 2006 ACT, Inc. is responsible for GMAT test development. A new automated essay scoring system is being used in conjunction with the ACT contract including IntelliMetric™ essay scoring.[113] Previously e-rater® has been used to score these exams from 1999 to 2005. Examples of where AES is used for summative evaluations include SAT, state testing, TOEFL, and PRAXIS.

For summative purposes, Dikli proposes "an effective way of using AES technology to score essays is to incorporate the AES system into the writing evaluation process as a second or third rater. As Monaghan and Bridgeman (2005) suggested, using an AES system as a check point to compare the scores assigned by human readers can be an effective way of incorporating the AES technology in writing assessment. In other words, the AES systems can be used both to verify human scoring and to represent a collection of human judges in large-scale writing assessments."[114]

AES FOR SPECIFIC POPULATIONS

Grimes and Warschauer (2006) used interviews, surveys, and classroom observations to study teachers and students using AES software to help revise their writing in five elementary and secondary schools. They report that in spite of generally positive attitudes toward AES as an aid for revision, the usage was lower than expected. Teachers scheduled little time for revising, and students used the information mainly to correct spelling errors. [115]

English Speakers and Speakers of Other Languages

At the post-secondary level Warschauer and Ware (2006) state that the ability to write well in English across diverse settings and for different audiences has become an imperative in second language education programs internationally as English becomes a global language. This has created demand for tools to help evaluate repeated drafts of student writing necessary in the teaching of second language writing. This is a shift in focus since the time prior to the Second World War, when much English language instruction focused on reading, as reading was the principal use of foreign languages in

112 Taylor, Alan R. (2006), pp. 19-20.
113 Rudner, L. M., Garcia, V., and Welch, C. (2006). An Evaluation of the IntelliMetricSM Essay Scoring System. *Journal of Technology, Learning, and Assessment* 4 (4), p. 4.
114 Dikli, Semire. (2006). An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning, and Assessment* 5 (1), pp. 27-28.
115 Grimes, D. & Warschauer, M. (2006). Automated Essay Scoring in the Classroom. p. 2.

many countries. Today as a result of the linguistic demands of a globalized informational society the educated population now needs to write in English for vocational or professional purposes.[116]

In discussing second language instruction, Ware and Warschauer (2006) state that computer mediated feedback is especially beneficial at the word, sentence, or paragraph level but "the dynamics of oral interaction allow for more free-flowing discussion, thereby result in more global changes to writing, such as general revokes of direction, purpose, or organization."[117]

Ware and Warschauer (2006) believe that notions about conventional forms of literacy and feedback are expanding as second language writers shift their audiences from single classroom teacher to peers and professionals across contexts. And that contrary to earlier speculation, the teacher's role in technology enhanced classrooms is not lessening but "rather, pedagogical framing and instructional guidance play a powerful role in shaping the success of online learning."[118]

Research related to the similar and different ways that native and ESL (English as Second Language) students and teachers could utilize AES engines in the K–12 classroom would be helpful to elementary and secondary schools. It would be interesting to note whether the feedback provided, especially to explain conventions, would need to be more detailed in terms of style, content and amount to those whose first language is not English.

Elliott states that "one of the best attributes of IntelliMetric™ is that it is capable of evaluating essay responses in multiple languages including English, Spanish, Hebrew, Bahasa, Dutch, French, Portuguese, German, Italian, Arabic, and Japanese (Elliot, 2003)"[119] which would help speakers whose first language is included in this list. A possible area of inquiry is whether English speakers learning one of the above languages would be helped by AES software and what adaptations are helpful to students learning languages with different alphabet systems. This question is posed as a result of a study conducted by Wolfe (2003)[120] where Test of English as a Foreign Language (TOEFL) candidates were given the choice of using paper and pencil or a word processing program to complete a direct writing assessment. Females, speakers of languages based on non-Roman/Cyrillic character systems, candidates from Africa and the Middle East, and those with less proficient English skills were more likely to choose handwriting than keyboarding. Younger candidates from Europe and older candidates from Asia were more likely to choose handwriting than their regional counterparts.

---

116 Warschauer, Mark and Ware, Paige. (06). Automated Writing Evaluation: Defining the Classroom Research Agenda. *Language Teaching Research* 10 (2), p. 1.

117 Ware, Paige, Warschauer, Mark, and (in press) (2006). Electronic Feedback and Second Language Writing, p. 19.

118 Ibid., p. 19.

119 Dikli, Semire. (2006). p. 18.

120 Wolfe, Edward W. (2003). Examinee Characteristics Associated With Choice of Composition Medium on the TOEFL Writing Section. *Journal of Technology, Learning, and Assessment* 2 (4), p. 1.

Students with Special Needs

Though not specific to AES, Ketterlin-Geller (2005) presents the steps for an application of universal design for assessment (UDA) intended to increase participation of students with disabilities and English-language learners in general education assessments by addressing student needs through customized testing platforms. She believes computer-based testing provides an optimal format for creating individually-tailored tests.[121] AES could have potential for assessing students with special needs.

Online Learning

New means of learning require new techniques. It seems appropriate to apply the techniques of automatic text evaluation to online discourse, which is central to much of distance education. It is now being used in some components of on-site education that were traditionally handled through face-to-face discussion and short written assignments. Online discourse has the advantage of already being in digital form, but it presents difficulties for AES scoring as (1) it is more free-form than even the loosest essay assignments, (2) the quantity of text produced by different students can vary greatly, (3) there are in problems of reference or deixis when the online discussion presupposes knowledge of shared experience taking place off-line, and (4) as online discourse occurs over time, how do you assess change? Some of the questions related to online discourse Bereiter proposes are: "Is the discussion getting anywhere? Is there evidence of learning? Are there changes in belief or interpretation? What happens to new ideas as they enter the discourse?"[122]

In this chapter the integral relationship between teaching, learning and assessment is outlined in relation to the teaching of writing and the assessment of student learning by means of essay and short answer responses. Attention is given to the role AES can play in both formative (assessment for learning) and summative (assessment of learning) assessment as well as how AES may be able to specifically address the learning needs of specific student populations in the K–12 community.

---

121 Ketterlin-Geller, Leanne R. (2005). Knowing What All Students Know: Procedures for Developing Universal Design for Assessment. *Journal of Technology, Learning, and Assessment* 4 (2), p. 1.

122 Bereiter, C. (2003). Foreword, p. ix.

6. Key Findings, Implications, and Recommendations

Computer-based assessment promises to both make obsolete many of the short comings of current high-stakes, state wide assessment systems and expand the capacity of such systems to measure rigorous standards in truly innovative ways.[123]

HIGHLIGHTS OF FINDINGS

Automated Essay Scoring (AES) is a relatively new field that elicits passionate responses from both its supporters and detractors, while being still under the radar for the great majority of educators in elementary and secondary schools in Canada. The purpose of this review was to: describe the technical state of the art in the field of scoring or grading of essays using computer technology; outline past and emergent practices and their pedagogical implications; summarize existing research; direct readers to more specific reference materials; provide a framework/blueprint for thinking about possible risks and potential in this field; and suggest directions for further research and development.

The amount and scope of AES literature has been increasing, especially since the turn-of-the-century (2000) which coincides with a greater expansion of AES engines being marketed commercially for use in the primary, secondary, and tertiary levels. AES is a young field that is growing exponentially as commercial programs for both formative and summative evaluation are marketed for an increasing variety of purposes.

Philosophically, teachers of writing, computational linguists and the computer scientists involved in Artificial Intelligence (AI) and other tools used in AES engines have very different viewpoints as to the advisability, reliability and validity of using AES. The National Council of Teachers of English (NCTE) position papers unequivocally state no writing should be scored by machines. Supporters of AES make many claims about the effectiveness and efficiency of AES in supporting writing instruction and for summative evaluation purposes.

The claims supporting the use of AES center around its perceived objectivity, potential for immediate feedback, cost efficiency, reduction of teacher marking time, assessment occurring in the same milieu

123 Rabinowitz, S. & Brandt, T. (2001). *Computer-Based Assessment: Can It Deliver on Its Promise?* p. 3. Retrieved from http://www.wested.org/cs/we/view/rs/568

as learning (computer versus paper and pencil) and more accessible statistical data to inform instruction. These factors according to its advocates lead to the ability to score more pieces of written work per student, thus leading to greater learning. The cautions presented by its detractors are its lack of human interaction, vulnerability to cheating, need for computer and software access, the restraints of prompts, and the need for a large corpus of sample text in order to train the system to each prompt.

While there has been research in this area, much of it has been conducted by firms which have proprietary interest in these programs. Research has focused mainly on the logistics of the AES programs, (i.e., How do AES scores compare to those of human raters? Is there validity and reliability?) While individual AES engines do provide scores comparable with human raters, there is a lack of a standard or unified measure to compare AES systems to each other to provide potential consumers with valid information for comparison.

For educators in the K–12 system there is scant research on the best pedagogical methods related to AES use to increase student learning. More research has been focused on college or university students and even at this level the focus is more on summative assessment for placement or program entry purposes.

Given the above findings, the following recommendations are presented to provide direction for future research and policy that will focus on the most beneficial manner to enhance K–12 instruction and assessment by investigating and implementing AES programs as appropriate.

RECOMMENDATIONS

Pedagogical Research

In considering whether AES is educationally beneficial to implement at the K–12 level in Canada, the following are examples of some of the issues to be addressed:

- examine the current literature and conduct further research as to the most effective and efficient ways to implement AES in order to increase K–12 student learning
- assess the economics of reducing the amount of time spent by teachers marking free text and essay responses, and therefore the educational efficiencies and effectiveness that might occur with redirection of this time into other educationally proven strategies to increase student achievement
- determine the professional development needs of teachers and other staff in order to use AES engines in ways that best promote student achievement and enhance teacher effectiveness and efficiency
- evaluate the cost of the software and hardware requirements for implementation and maintenance of AES programs for both formative and summative assessment
- gather feedback from students of their perceptions of the effectiveness of the AES programs in (1) formative evaluation situations and (2) summative evaluation situations
- conduct research into the apparent reluctance by students to perform multiple edits when using formative AES engines
- compare differences in using AES engines for writing instruction by native English speakers and ESL students

Recommendation #1

Expand research to focus more on the potential uses and effectiveness of AES in the K–12 setting for the instruction and assessment of writing and the assessment of other learning through essay and short answer responses. Independent researchers could investigate both the formative and summative applications of AES at both the classroom and the large-scale levels of instruction and assessment for different student populations in K–12.

Recommendation #2

Research whether AES scoring engines are effective in the following types of scenarios when AES is used for summative evaluation purposes: (1) determining K–12 students' placement in levels for writing instruction; (2) determining a Grade 12 student's future success when selecting potential students for entry into college or university programs; and, (3) determining future success when AES is used as part of summative evaluations that select people for entry into a specific profession, e.g. teaching.

Future Directions for Research on AES Technology

Valenti et al. (2003) state that "The most common problems encountered in the research on automated essay grading are the absence both of a good standard to calibrate human marks and of a clear set of rules for selecting master texts. A first conclusion obtained is that in order to really compare the performance of the systems some sort of unified measure should be defined. Furthermore, the lack of standard data collection is identified. Both these problems represent interesting issues for further research in this field."[124] AES should strive to score even more effectively than human raters.

In ongoing assessment programs that evaluate writing, care is taken to ensure that scoring of essays is consistent from one administration to another. Often sets of essays from a previous administration are rescored by markers in the current session so that marker drift, if any, can be estimated.

Reliable and efficient human holistic scoring is based on a scale with a limited number of categories: typically six. Training of the AES scoring engine amounts to selecting regression variables and establishing their relative weights and/or generating the probabilities that certain characteristics individually or jointly predict the categorical scores assigned to a set of essays by a panel of experts. In the scoring phase, regression scores or probability distributions are generated which closely approach a continuous scale. These scores are then collapsed to the original six categories thus losing formation.

Recommendation #3

Research should be done on whether AES programs could be enhanced so that their scores more closely approximate a continuous scale. Currently, a six-point scale is used by both humans and computer programs. Any further studies correlating AES scores to other writing constructs should compare the more continuous scales with the six-point categorical scales that are currently used.

---

124 Valenti, Salvatore, Nitko, Anthony J., and Cucchiarelli, Alessandro. (03). An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education* 2 (Information Technology for Assessing Student Learning Special Issue), p. 319.

Recommendation #4

A similar test should be applied to AES scoring. If an AES engine is trained on a set of essays on one topic and applied to a second topic, how well would those scores correlate with the scores of the same engine trained for the second topic?

Recommendation #5

Research should be conducted to examine the equivalency of AES scores produced from rubrics used in different jurisdictions so that a variety of prompts can be used across educational systems without AES engines having to be retrained for each specific prompt.

Recommendation #6

Research should be done on whether AES systems can generate more detailed descriptive feedback for different prompts so that learners do not need to respond to a restricted set of prompts but would be able to gain feedback and have their written work assessed on any topic that they wished.

Recommendation #7

Develop a unified measure in order to compare the relative performance of different AES programs so that accurate comparisons can be made to allow the selection of the best tools to enhance K–12 student learning.

Future Directions for Policy

There is a large and growing body of research on language learning, use, and assessment that must be used to improve assessment on a systematic and regular basis . . . Anyone charged with the responsibility of designing an assessment program must be cognizant of this body of research and must stay abreast of developments in the field. Thus, assessment programs must always be under review and subject to change by well-informed faculty, administrators, and legislators.[125]

Policy recommendations need to be considered at the school, district, provincial and national level. Instead of not addressing the issues related to AES and by default allowing the commercial vendors to set the agenda for implementation of it in K–12 settings in Canada, a proactive role needs to be taken in order to ensure that the adoption of AES, if it occurs, is as educationally effective and efficient as possible and that all stakeholders have participated in the discussion. Dikli suggests that it would be interesting to envision AES systems as a free public utility rather than proprietary, vendor-created and owned. This would allow more teachers and students to benefit from AES systems in writing classrooms.[126]

Recommendation #8

Establish a joint task force of professionals from the Canadian K–12 education teaching, assessment and policymaking communities to examine AES and its related issues to make recommendations as to its educational appropriateness for Canadian K–12 learners.

---

125 Conference on College Composition and Communication (1995), pp. 4-5.
126 Dikli, Semire. (2006). p. 27.

SUMMARY

AES has the potential for being an exciting innovation that could revolutionize learning and assessment in many curriculum areas at all levels of instruction. According to its critics, however, it has the possibility to dehumanize and mechanize writing instruction and assessment or even harm writing instruction. Further research and educational dialogue are necessary for educators in Canada and elsewhere in the world to determine the educational significance and appropriate role for AES in the education of K–12 students.

GLOSSARY

The following are terms used to describe the process of grading essays or written work using computers:

AES automated essay scoring
AEG automated essay grading
AWE automated writing evaluation
computer essay grading
computer graded essays
computerized essay scoring
computerized essay assessment
machine scoring of essays

The following are some of the AES models mentioned in Chapter 3:

Bayesian Essay Test Scoring System™ (BETSY)
Critique and Criterion℠ (Educational Testing Service - ETS)
E-rater® (Educational Testing Service - ETS)
Intelligent Essay Assessor™ (IEA) (Pearson Knowledge Technologies -PKT)
IntelliMetric™ (Vantage Learning)
MarkIT
MY Access!® (Vantage Learning)
Project Essay Grader™ (PEG)

APPENDIX A

CCCC Position Statement on Teaching, Learning, and Assessing Writing in Digital Environments

Conference on College Composition and Communication - February 2004

[In the spring of 2003, then-Chair of CCCC Shirley Wilson Logan appointed a CCCC Committee whose purpose was to create a position statement governing the teaching, learning, and assessing of writing in digital environments. This is the document this group produced; it was adopted by the CCCC Executive Committee as of February 25, 2004.]

Submitted by the CCCC Committee on Teaching, Learning, and Assessing Writing in Digital Environments (Kathleen Yancey, Chair; Andrea Lunsford; James McDonald; Charles Moran; Michael Neal; Chet Pryor; Duane Roen; Cindy Selfe)

Increasingly, classes and programs in writing require that students compose digitally. Such writing occurs both in conventional "face-to-face" classrooms and in classes and programs that are delivered at a distance. The expression "composing digitally" can refer to a myriad of practices. In its simplest form, such writing can refer to a "mixed media" writing practice, the kind that occurs when students compose at a computer screen, using a word processor, so that they can submit the writing in print (Moran). Such writing may not utilize the formatting conventions such as italics and bold facing available on a word processor; alternatively, such writing often includes sophisticated formatting as well as hyper textual links. Digital composing can take many other forms as well. For example, such composing can mean participating in an online discussion through a listserv or bulletin board (Huot and Takayoshi). It can refer to creating compositions in presentation software. It can refer to participating in chat rooms or creating web pages. It can mean creating a digital portfolio with audio and video files as well as scanned print writings. Most recently, it can mean composing on a class weblog or wiki. And more generally, as composers use digital technology to create new genres, we can expect the variety of digital compositions to continue proliferating.

The focus of writing instruction is expanding: the curriculum of composition is widening to include not one but two literacies: a literacy of print and a literacy of the screen. In addition, work in one medium is used to enhance learning in the other.

As we refine current practices and invent new ones for digital literacy, we need to assure that principles of good practice governing these new activities are clearly articulated.

Assumptions

Courses that engage students in writing digitally may have many features, but all of them should :

(a) introduce students to the epistemic (knowledge-constructing) characteristics of information technology, some of which are generic to information technology and some of which are specific to the fields in which the information technology is used;
(b) provide students with opportunities to apply digital technologies to solve substantial problems

common to the academic, professional, civic, and/or personal realm of their lives;
(c) include much hands-on use of technologies;
(d) engage students in the critical evaluation of information (see American Library Association, "Information Literacy"); and
(e) prepare students to be reflective practitioners.

As with all teaching and learning, the foundation for teaching writing digitally must be university, college, department, program, and course learning goals or outcomes. These outcomes should reflect current knowledge in the field (such as those articulated in the "WPA Outcomes Statement"), as well as the needs of students, who will be expected to write for a variety of purposes in the academic, professional, civic, and personal arenas of life. Once programs and faculty have established learning outcomes, they then can make thoughtful decisions about curriculum, pedagogy, and assessment.

Writing instruction is delivered contextually. Therefore, institutional mission statements should also inform decisions about teaching writing digitally in the same ways that they should inform any curricular and pedagogical decisions.

Regardless of the medium in which writers choose to work, all writing is social; accordingly, response to and evaluation of writing are human activities, and in the classroom, their primary purpose is to enhance learning.

Therefore, faculty will:

(a) incorporate principles of best practices in teaching and learning. As Chickering and Ehrmann explain, those principles are equally applicable to face-to-face, hybrid, and online instruction
    - •Good Practice Encourages Contacts Between Student and Faculty
    - •Good Practice Develops Reciprocity and Cooperation Among Students
    - •Good Practice Uses Active Learning Techniques
    - •Good Practice Gives Prompt Feedback
    - •Good Practice Emphasizes Time on Task
    - •Good Practice Communicates High Expectations
    - •Good Practice Respects Diverse Talents and Ways of Learning

(b) provide for the needs of students who are place-bound and time-bound.

(c) be guided by the principles outlined in the CCCC "Writing Assessment: A Position Statement" for assessment of student work in all learning environments—in face-to-face, in hybrid, and in online situations. Given new genres, assessment may require new criteria: the attributes of a hyper textual essay are likely to vary from those of a print essay; the attributes of a weblog differ from those of a print journal (Yancey). Because digital environments make sharing work especially convenient, we would expect to find considerable human interaction around texts; through such interaction, students learn that humans write to other humans for specific purposes. Good assessment requires human readers.

Administrators with responsibilities for writing programs will:

(a) assure that all matriculated students have sufficient access to the requisite technology, thus bridg-

ing the "digital divide" in the local context. Students who face special economic and cultural hurdles (see Digital Divide Network) as well as those with disabilities will receive the support necessary for them to succeed;

(b) assure that students off campus, particularly in distance learning situations, have access to the same library resources available to other students (see American Library Association, "Guidelines for Distance Learning");

(c) assure that reward structures for faculty teaching digital writing value such work appropriately. Department, college, and institutional policies and procedures for annual reviews and for promotion and tenure should acknowledge the time and intellectual energy required to teach writing digitally (see CCCC "Promotion and Tenure" and "Tenure and Promotion Cases for Composition Faculty Who Work with Technology"). This work is located within a new field of expertise and should be both supported—with hardware and software—and recognized. Similarly, institutions that expect faculty to write for publication must have policies that value scholarly work focused on writing in digital environments—the scholarship of discovery, application/engagement, integration, and teaching (see Boyer; Glassick, Huber, and Maeroff; Shulman);

(d) assure that faculty have ready access to diverse forms of technical and pedagogical professional development before and while they teach in digital environments. Such support should include regular and just-in-time workshops, courses, individual consultations, and Web resources;

(e) provide adequate infrastructure for teaching writing in digital environments, including routine access to current hardware; and

(f) develop equitable policies for ownership of intellectual property that take effect before online classes commence

Writing Programs, in concert with their institutions, will:

(a) assess students' readiness to succeed in learning to write in digital environments. Programs should assess students' access to hardware, software and access tools used in the course, as well as students' previous experience with those tools. In order to enhance learning, programs may also assess students' attitudes about learning in online environments; and

(b) facilitate the development of electronic portfolios where such programs are in place or are under consideration. As important, writing programs will work to help develop the infrastructure and the pedagogy to assist students in moving their portfolios from one course to another, one program to another, one institution to another, as well as from educational institutions to the workplace, working to keep learning at the center of the enterprise and to assure that students learn to use the technology, not just consume it. To accomplish this goal, institutions need to work with professional organizations and software manufacturers to develop portfolio models that serve learning.

A Current Challenge: Electronic Rating

Because all writing is social, all writing should have human readers, regardless of the purpose of the writing. Assessment of writing that is scored by human readers can take time; machine-reading of placement writing gives quick, almost-instantaneous scoring and thus helps provide the kind of quick assessment that helps facilitate college orientation and registration procedures as well as exit assessments.

The speed of machine-scoring is offset by a number of disadvantages. Writing-to-a-machine violates the essentially social nature of writing: we write to others for social purposes. If a student's first writing-experience at an institution is writing to a machine, for instance, this sends a message: writing at this institution is not valued as human communication—and this in turn reduces the validity of the

assessment. Further, since we can not know the criteria by which the computer scores the writing, we can not know whether particular kinds of bias may have been built into the scoring. And finally, if high schools see themselves as preparing students for college writing, and if college writing becomes to any degree machine-scored, high schools will begin to prepare their students to write for machines.

We understand that machine-scoring programs are under consideration not just for the scoring of placement tests, but for responding to student writing in writing centers and as exit tests. We oppose the use of machine-scored writing in the assessment of writing.

REFERENCES

Alberta Teachers' Association (2002). Alberta Teachers' Association: Ongoing Issues - Hours of Instruction. pp.1-2.Retrieved from http://www.teacher.ab.ca/Issues+In+Education/Ongoing+Issues/Hours+of+Instruction.htm

Assessment and Testing Study Group of the NCTE Executive Committee. (2004). Framing Statements on Assessment: Revised Report of the Assessment and Testing Study Group of the NCTE Executive Committee. pp.1-4.Retrieved from http://www.ncte.org/about/over/positions/category/assess/118875.htm

Attali, Y. & Burstein, J. (2006, February). Automated Essay Scoring With E-Rater® V.2. *Journal of Technology, Learning, and Assessment* 4(3), pp.1-31.Retrieved from http://escholarship.bc.edu/jtla/vol4/3/

Bennett, R. E. (2001, February). How the Internet Will Help Large-Scale Assessment Reinvent Itself. *Education Policy Analysis* 9(5), pp.1-39.Retrieved from http://epaa.asu.edu/epaa/v9n5.html

Berard, Y. & Cromer Brock, K. (2006, July). Who's Keeping Score? Grading the TAKS Essay. *Star- Telegram*. Fort Worth, Texas. Retrieved from http://www.dfw.com/mld/dfw/community/15105085.htm

Bereiter, C. (2003). Foreword. In *Automated Essay Scoring: A Cross-Disciplinary Perspective* ed. by Shermis, Mark D. and Burstein, Jill, pp. vii-ix.

Broad, Bob. (2006). More Work for Teacher? Possible Futures of Teaching Writing in the Age of Computerized Writing Assessment. In *Machine Scoring of Student Essays* ed. by Ericsson, Patricia and Haswell, Richard, pp. 221-233.

Burstein, J. (2003). The E-rater® Scoring Engine: Automated Essay Scoring with Natural Language Processing. In *Automated Essay Scoring: A Cross-Disciplinary Perspective* ed. by Shermis, Mark D. and Burstein, Jill, pp. 113-122.

Conference on College Composition and Communication (1995). Writing Assessment: A Position Statement. pp.1-9. Retrieved from http://www.ncte.org/cccc/resources/positions/123784.htm

Conference on College Composition and Communication (2004). *CCCC Position Statement on Teaching, Learning, and Assessing Writing in Digital Environments*. p.1. Retrieved from http://www.ncte.org/cccc/resources/positions/123773.htm

Dikli, S. (2006, August). An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning, and Assessment* 5(1), pp.1-35. Retrieved from http://escholarship.bc.edu/jtla/vol5/1/

Elliott, Scott. (2003). Intellimetric™: From Here to Validity. In *Automated Essay Scoring: A Cross-Disciplinary Perspective* ed. by Shermis, Mark D. and Burstein, Jill, pp. 71-86.

Ericsson, Patricia and Haswell, Richard. (2006). Introduction. In *Machine Scoring of Student*

Essays ed. by Ericsson, Patricia and Haswell, Richard, pp. 1-268.

Ericsson, Patricia and Haswell, Richard, eds. (2006). *Machine Scoring of Student Essays*. Logan, Utah: Utah State University Press, pp. 1-268.

Freitz, E. (2006). Book Review: Machine Scoring of Student Essays: Truth and Consequences. *The CEA Forum* 35(1), pp.1-2. College English Association. Retrieved from http://www2.widener.edu/~cea/351fleitz.htm

Goldberg, A., Russell, M., & Cook, A. (2003, February). The Effect of Computers on Student Writing: A Meta-Analysis of Studies From 1992 to 2002. *Journal of Technology, Learning, and Assessment* 2(1), pp.1-51. Retrieved from http://escholarship.bc.edu/jtla/vol2/1/

Grimes, D. & Warschauer, M. (2006). *Automated Essay Scoring in the Classroom.* pp.1-30. Retrieved from www.gse.uci.edu/faculty/markw/aera-2006-aes.pdf

Haladyna, T. M. & Olsen, R. M. (2006, June). *Threats to Validity in Large-Scale Writing Performance Tests: What Are These Threats and What Should Be Done About It?* pp.1-25. Retrieved from www.sanjuan.edu/accountability/program-evaluations/documents/Validity-v4.0.pdf

Haswell, Richard. (2006). Automatons and Automated Scoring. In *Machine Scoring of Student Essays* ed. by Ericsson, Patricia and Haswell, Richard, pp. 57-78.

Keith, Timothy Z. (2003).Validity and Automated Essay Scoring Systems. In *Automated Essay Scoring: A Cross-Disciplinary Perspective* ed. by Shermis, Mark D. and Burstein, Jill, pp. 147-168.

Ketterlin-Geller, L. R. (2005, November). Knowing What All Students Know: Procedures for Developing Universal Design for Assessment. *Journal of Technology, Learning, and Assessment* 4(2), pp.1-22. Retrieved from http://escholarship.bc.edu/jtla/vol4/2/

Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The Debate on Automated Essay Grading: The Intelligent Essay Assessor. *IEEE* pp. 27-31. Retrieved from http://www.pearsonkt.com/papers/IEEEdebate2000.pdf

McAllister, Ken S. and White, Edward M. (2006). Interested Complicities. In *Machine Scoring of Student Essays* ed. by Ericsson, Patricia and Haswell, Richard, pp. 8-27.

Mesthene, E. G. (1969). Some General Implications of the Research of the Harvard University Program on Technology and Society. *Technology and Culture* 10(4), pp. 489-513.

Naylor, C. & Malcolmson, J. (2001, September). BCTF Research Report 2001-WLC-02 *"I Love Teaching English, but...." A Study of the Workload of English Teachers in B.C. Secondary Grades.* RT01-0036 pp.1-48.: BCTF. Retrieved from www.bctf.ca/ResearchReports/2001wlc02

NCTE Task Force on SAT and ACT Writing Tests (2005, April). *The Impact of the SAT and ACT Timed Writing Tests: Report From the NCTE Task Force on SAT and ACT Writing Tests*. pp.1-15. Retrieved from http://www.ncte.org/library/files/About_NCTE/Press_Center/SAT/SAT-ACT-tf-report.pdf

Page, Ellis Batten. (2003).Project Essay Grade: PEG. In *Automated Essay Scoring: A Cross-Disciplinary Perspective* ed. by Shermis, Mark D. and Burstein, Jill, pp. 43-54.

Pearson Knowledge Technologies WriteToLearn. pp. 1-4. Retrieved from www.pearsonkt.com/prodWTL.shmtl

Ponisciak, Steve and Johnson, Valen. (2003). Bayesian Analysis of Essay Grading. In *Automated Essay Scoring: A Cross-Disciplinary Perspective* ed. by Shermis, Mark D. and Burstein, Jill, pp. 181-192.

Rabinowitz, S. & Brandt, T. (2001). *Computer-Based Assessment: Can It Deliver on Its Promise?* pp.1-8. San Francisco, California: WestED®. Retrieved from http://www.wested.org/cs/we/view/rs/568

Ridgway, J., McCusker, S., & Pead, D. (2004). *Literature Review of E-Assessment*. 10, pp.1-47. Bristol, United Kingdom: Futurelab. Retrieved from http://www.futurelab.org.uk/research/lit_reviews.htm#lr10

Riedel, E., Dexter, S., Scharber, C., & Doering, A. (2005, April). Experimental Evidence on the Effectiveness of Automated Essay Scoring in Teacher Education Cases. *86th Annual Meeting of the American Educational Research Association* pp.1-16. Retrieved from http://www.etips.info/papers/2experimental.html

Ripley, M. (2004, April). E-Assessment: An Overview. *QCA keynote speech.*

Rudner, L. M., Garcia, V., & Welch, C. (2006, March). An Evaluation of the IntelliMetric[SM] Essay Scoring System. *Journal of Technology, Learning, and Assessment* 4(4), pp.1-21. Retrieved from http://escholarship.bc.edu/jtla/vol4/4/

Rudner, L. & Gagne, P. (2001). An Overview of Three Approaches to Scoring Written Essays by Computer. *Practical Assessment, Research & Evaluation* 7(26). Retrieved from http://pareonline.net/getvn.asp?v=7&n=26

Russo, A. (2002, April). Mixing Technology and Testing. *School Administrator* 4(59), pp.6-12. American Association of School Administrators. Retrieved from http://www.aasa.org/publications/saarticledetail.cfm?ItemNumber=2601&snItemNumber=950&tnItemNumber=951

Sarah Schmidt (2006, September). Most Undergrads Admit to Cheating, Study Finds. *Times Colonist* pp.A2. Victoria, BC

Scalise, K. & Gifford, B. (2006, June). Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms. *Journal of Technology, Learning, and Assessment* 4(6), pp.1-44.Retrieved from http://escholarship.bc.edu/jtla/vol4/6/

Shermis, Mark D. and Burstein, Jill, eds. (2003). *Automated Essay Scoring: A Cross-Disciplinary Perspective.* Mahwah, New Jersey: Lawrence Erlbaum Associates, pp. 1-238.

Shermis, Mark D. and Burstein, Jill. (2003). Preface. In *Automated Essay Scoring: A Cross-Disciplinary Perspective* ed. by Shermis, Mark D. and Burstein, Jill, pp. xi-xii.

Taylor, A. R. (2006, May). *A Future in the Process of Arrival: Using Computer Technologies for the Assessment of Student Learning.* 22 pp.1-114.: SAEE. Retrieved from http://www.tasainstitute.com/029.pdf

Valenti, S., Nitko, A. J., & Cucchiarelli, A. (2003). An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education* 2 (Information Technology for Assessing Student Learning Special Issue), pp. 319-30.

Vantage Learning (2006, August). MY Access®. Vantage Learning. Retrieved from http://www.vantagelearning.com/myaccess/

Ware, P., Warschauer, M., and (in press). (2006). Electronic Feedback and Second Language Writing. In *Feedback and second language writing* ed. by Hyland, K. and Hyland, F.

Warschauer, M. & Ware, P. (2006). Automated Writing Evaluation: Defining the Classroom Research Agenda. *Language Teaching Research* 10(2), pp.1-24. Retrieved from http://www.gse.uci.edu/faculty/markw/awe.pdf

Williams, R. & Dreher, H. (2005). Formative Assessment Visual Feedback in Computer Graded Essays. *The Journal of Issues in Informing Science and Information Technology* 2 pp. 23-32.

Wolfe, E. W. (2003, December). Examinee Characteristics Associated With Choice of Composition Medium on the TOEFL Writing Section. *Journal of Technology, Learning, and Assessment* 2(4), pp.1-25. Retrieved from http://escholarship.bc.edu/jtla/vol2/4/

Writing Study Group of the NCTE Executive Committee (2004). *NCTE Beliefs About the Teaching of Writing*. pp.1-15. Retrieved from http://www.ncte.org/about/over/positions/category/write/118876.htm

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2005, July). Evaluating Computer Automated Scoring: Issues, Methods, and an Empirical Illustration. *Association of Test Publishers Journal* July 2005 pp.1-42. Retrieved from http://www.testpublishers.org/journal.htm

Automated Essay Scoring: A Literature Review

Susan M. Phillips

Automated Essay Scoring (AES) is an emerging field that has strong supporters and equally strong detractors. Whether it is viewed as a promising tool to improve the potential of student writing or a threat that removes the teacher from the evaluation process, AES is an issue on the leading edge of assessment for K-12. This literature review covers the disciplinary approaches of writing instruction, computational linguistics, and computer science, and the spectrum of perspectives that derive from them.

Readers will be introduced to a variety of AES models, practical issues, recent research and future directions. Analytical tools and classification systems such as Bayesian text classification, latent semantic analysis and natural language processing are presented as well as a variety of AES engines currently on the market that offer both summative and formative evaluation programs.

Susan M. Phillips provides an analysis of pedagogical issues and controversies that arise from the adoption of AES with a focus on best practices to enhance student achievement. Key findings, implications and recommendations for researchers, educators and policy makers are based on a thorough understanding of the current research and lay the foundation for the thoughtful use of AES in K-12 schools in Canada.

This work was commissioned through the Technology Assisted Student Assessment (TASA) Institute established by SAEE in September 2004 for the study of technology's role in the K-12 domain. For more information see www.tasainstitute.com

SOCIETY FOR THE ADVANCEMENT OF EXCELLENCE IN EDUCATION

225 - 1889 Springfield Road, Kelowna BC V1Y 5V5

Telephone 250.717.1163 : Fax 250.717.1134

Email info@saee.ca, Website: http://www.saee.ca