Council of Writing Program Administrators

WPA-CompPile Research Bibliographies

Norbert Elliot; Anne Ruggles Gere; Gail Gibson; Christie Toth; Carl Whithaus; Amanda Presswood

Uses and Limitations of Automated Writing Evaluation Software

(WPA-CompPile Research Bibliographies, No. 23)

December 2013

This annotated bibliography is intended to provide Writing Program Administrators (WPAs) with an overview of research in Automated Writing Evaluation (AWE). The bibliography provides a system of validation, reviews the most recent research studies according to that system, and in doing so identifies the potentials for and limitations of AWE software. *Automated Writing Evaluation* is one of several terms we could have chosen to describe our focus. "Automated Essay Scoring (AES)," a term often used by researchers in the field, is insufficiently capacious because "essay" does not include the various genres of student writing; similarly, "scoring" connotes only summative evaluation, and AWE can provide formative evaluation. For similar reasons, we rejected the terms "Automated Essay Evaluation (AEE)" and "machine scoring."

Research on AWE is moving quickly, and the citations included in this bibliography may soon be outdated. Nevertheless, this introduction offers enduring explanations and principles designed to help WPAs consider any AWE system they encounter in the near future. One term currently used in many discussions of AWE is "writing construct," which refers to the way writing is understood by a given community. The *Framework for Success in Postsecondary Writing* (2011), for example, emphasizes rhetorical knowledge, critical thinking, flexible writing processes, and ability to compose in multiple environments. A given AWE system might have the capacity to identify only some features of the writing construct, such as defined textual features, the presence of discourse elements, and word choice. Much of the current controversy surrounding AWE, controversy typified by the NCTE Task Force on Writing Assessment (2013) position statement "Machine Scoring Fails the Test," stems from conflicting views about what cognitive, intrapersonal, and interpersonal domains of writing are represented in any given instance of assessment.

A second set of potentially unfamiliar terms, drawn from computational linguistics, includes *Natural Language Processing* (NLP), which describes AWE systems that measure such features as syntactic construction, appropriate vocabulary use, and knowledge of conventions, and *Latent Semantic Analysis* (LSA), where corpus-based statistical modeling is used to analyze writing in terms of vocabulary usage. These algorithmic models allow machines to produce scores or evaluative comments on measurable aspects of writing, but—as acknowledged by those who design and use them—they do not "read" student writing as would a human. So, in their present state, these systems cannot judge critical thinking, rhetorical knowledge, or the ability of a writer to adapt to a given audience.

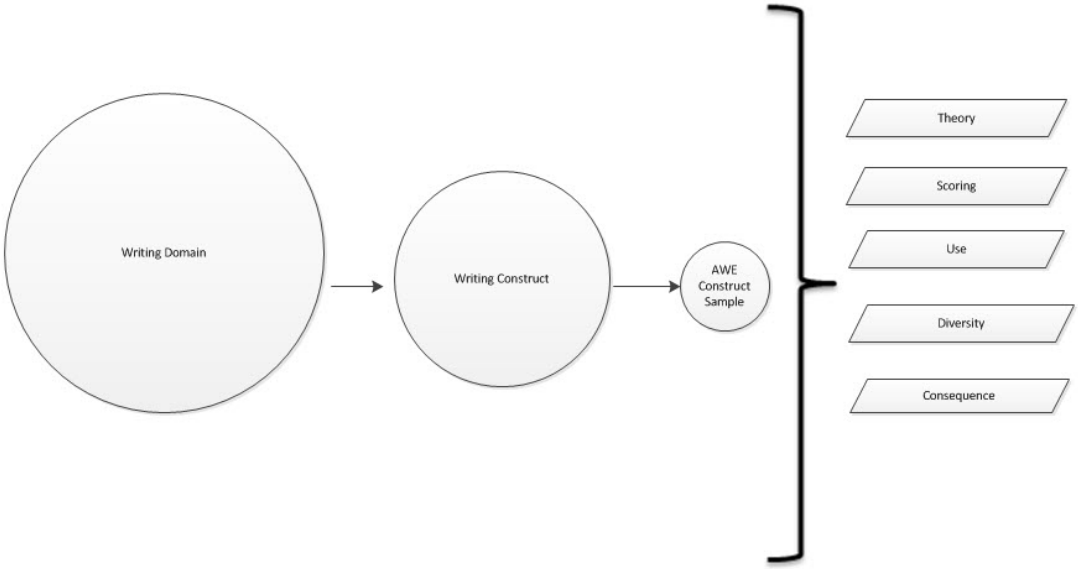
*Cite as: Norbert Elliot; Anne Ruggles Gere; Gail Gibson; Christie Toth; Carl Whithaus; Amanda Presswood. (December 2013). Uses and Limitations of Automated Writing Evaluation Software, WPA-CompPile Research Bibliographies, No. 23. *WPA-CompPile Research Bibliographies*. http://comppile.org/wpa/bibliographies/Bib23/AutoWritingEvaluation.pdf. Date of access.

One of the principles WPAs consider in any form of assessment is the impact of the assessment itself, and discussions of the consequences of AWE system use take on particular urgency when considering diverse student populations. To date, there has been relatively little research that disaggregates AWE outcomes by sub-groups. However, the handful of studies tackling aspects of this question suggest that AWE may have disparate impacts based on gender, ethnicity, nationality, and native language, privileging or penalizing some cultural backgrounds and languages over others. Questions about disparate impact are particularly urgent given the widespread adoption of AWE at precisely the kinds of institutions that serve a disproportionate number of students from diverse ethnic, linguistic, and socioeconomic backgrounds.

The intersections of construct representation, computational capability, and impact presently fuel lively discussion in the popular press, within professional associations, and among individual researchers. It is not our aim to enter this debate. Rather, our goal is to put into the hands of WPAs conceptual tools and empirical evidence for making their own decisions in local contexts. The emphasis on local use is critically important. Validity is not a unitary concept, but rather a process of gathering information to justify the use of any assessment. As such, WPAs can benefit from both frameworks and peer-reviewed studies that provide guidance in making judgments regarding the use of all types of writing assessment, including AWE. The process of validating any particular AWE software within a given context can include the possibility of rejecting that software for that use—i.e., a validation study may lead a WPA to conclude that a piece of software does not support a validity argument for its use in the local assessment the WPA is conducting.

One of the challenges facing WPAs is determining whether the construct of writing elicited by a given AWE aligns with the construct embedded in their writing program's curriculum. Validation, or the process of determining what evidence must be gathered to determine alignment and justify use, can be accomplished by organizing and analyzing information about a given system using categories of evidence. In Figure 1, we offer a model by which WPAs may validate AWE system use—or reject that use—based on the validation framework of Kane (2006, 2013).

Figure 1. A Conceptual System for Validation of AWE Use in Local Settings



Moving from left to right, the model begins with the broadest conceivable domain: all instances of writing genres in all imaginable settings. The middle circle refers to the construct of writing used by the WPA's specific institutional site, and the third circle refers to the construct of that given AWE system uses to assess samples of writing. Tracing the writing construct from this broadest domain to a specific assessment instance offers a method of identifying types of evidence that may be used to justify, or argue against, AWE use in local settings.

The analytic model suggested in Figure 1 can also be used to categorize a wide variety of research about AWE systems. We have therefore used the five categories of validity evidence from Figure 1 to organize the bibliography. While these categories of evidence are themselves limited, they can provide a systematic way for WPAs to engage in the process of AWE validation. To assist with this process, we offer two additional tables: an overview of AWE systems and products and a set of questions that can serve as an initial heuristic for local validation of the use of AWE.

Table 1. Overview of Automated Essay Scoring (AES), Automated Essay Evaluation (AEE), and Automated Writing Evaluation (AWE) Products

This table is intended to offer Writing Program Administrators basic information about the major systems currently in the market. This field is swiftly expanding and changing, but the information here is current as of October 2013. This table supplements the CompPile Annotated Bibliography on automated writing evaluation.

System	Developer	Platform	Scoring [1]	Primary Use	Primary Market	Demonstrations
ACCUPLACER®	College Board	MyFoundationsLab®		Course placement	Higher education	No online demo. Sample questions available here
AutoScore	American Institutes for Research	N/A	NLP	Assessment	K-12 market	Product overview here (demo page in development)
Mosaic™	TB McGraw-Hill	Writing Roadmap™	NLP	Writing instruction and assessment	Grades 3-12	Video overview of product
CRASE™	Pacific Metrics	N/A	NLP	Large-scale assessment	K-12 market	Brochure product overview
e-rater®	(ETS) Educational Testing Service	Criterion™	NLP	Writing instruction and assessment	K-12 and higher education markets	Video overview of product
Intelligent Essay Assessor (IEA)™	Pearson Knowledge Technologies	Write to Learn!	LSA	Writing instruction and assessment	K-12 market	Overview of product designed for student and teacher review
Intellimetric®	Vantage Learning McCann/Cengage Learning	MY Access!® Write Experience	NLP	Writing instruction and assessment	K-12 (MyAccess) and higher education (Write Experience)	Link to Intellimetric demo Link to Write Experience overview
Lexile® Writing Analyzer	MetaMetrics	Writing Analyzer	NLP	Writing instruction and assessment	K-12 market	Brief overview available here
LightSIDE	LightSide Labs	In development	Machine learning	Writing instruction, assessment, and peer review	K-12 market emphasized	Video overview of product
Project Essay Grade (PEG)™	Measurement Incorporated	N/A	Statistical techniques	Writing practice and assessment	Company notes partner relationships in K-12 and higher education markets	No online demo available

[1] The abbreviation NLP refers to Natural Language Processing; the abbreviation LSA refers to Latent Semantic Analysis. These terms are defined more fully in the companion bibliography. For further discussion, see Shermis, M. D., Burstein, J., & Bursky, S. A. (2013). Introduction to automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp. 2-15). New York, NY: Routledge.

Table 2. Sample Validity Questions for Writing Program Administrators: Evaluation of AWE Systems

CATEGORY OF VALIDATION EVIDENCE	Alignment of Writing Program *Theory*	Relationship to Institutional *Scoring* Practices	Practical *Use* of Assessment Results	Institutional *Diversity* Frameworks	Intended and Unintended *Consequence* of AWE Use within the Writing Program
Sample Question	What theory of writing has your writing program adopted? How does the AWE under consideration align with that theory of writing?	What are the scoring and evaluation practices used in your writing program? How does the AWE under consideration support those practices?	How are assessment results presently used in your writing program? How will the proposed AWE system influence that use?	In terms of equity and fairness, how is student diversity supported in your writing program? How does the proposed AWE strengthen your writing program's dedication to diversity?	How does your writing program deal with negative and positive impacts of the assessment on various communities? How will the proposed AWE decrease negative impact and support positive impact?

We turn now to the annotations themselves.

Section 1: Theory

In many ways, the debate over the growing use of Automated Writing Evaluation (AWE) systems in high-stakes testing and in classroom practice reflects a theoretical divide in understandings of language and writing. Computational methods of assessing writing rely on cognitive and psychological models of language processing that can be at odds with theoretical understandings of writing as a rhetorically complex and socially embedded process that varies based on context and audience. Research presented here addresses issues raised by theories or constructs of writing and the ways that automated assessment technologies either align or fail to align with those theories.

Condon, William

Large-scale assessment, locally-developed measures, and automated scoring of essays

Assessing Writing 18.1 (2013), 100-108

> This article suggests that the debate over the use of machines to score large-scale evaluations of student writing overlooks a more fundamental concern about the inability of such tests—however scored—to accurately reflect student abilities and instructional needs. By design, the types of tests that can be scored by automated systems generate only short writing samples produced under tight time restrictions. Because of this, Condon argues that such tests do not fully reflect widely accepted writing constructs and are "poor predictors of students' success in courses that require them to think, to write with an awareness of purpose

and audience, and to control the writing process" (p. 103). Condon acknowledges that computers are able to analyze some syntactical aspects and count certain features of writing (number of words, for instance, or average sentence length), and he suggests that such evaluation can offer a tool for writing instruction focused on specific textual features. But such use is limited, the author argues, and for now, overshadowed by increasing reliance on automated scoring engines in high-stakes assessments. Condon argues that such tests should not be used to assess student writing, whether in the context of admissions and placement or in the context of formative and summative measures of student achievement. Instead, writing instructors should look to richer forms of assessment, including course performance and portfolio evaluation, that reflect institutional context and provide a more complete measure of student understanding of the full writing construct. The article draws specific attention to the use of automated scoring systems for remediation placement, arguing that inaccurate placements into remedial coursework can limit student growth and discourage college persistence.

KEYWORDS: machine-scoring, critique, construct-validity, alternative, assessment, authentic, model, reconceptualization, learning-theory, large-scale, rhetorical, production, social, cultural

Deane, Paul

On the relation between automated essay scoring and modern views of the writing construct

Assessing Writing 18.1 (2013), 7-24

This article argues that future uses for Automated Writing Evaluation (AWE) should be embedded in an understanding of writing not only as a social and cultural process but also as requiring specific cognitive skills that machine analysis can help to develop. The author focuses on theoretical discussion of ETS's e-rater® program. This article acknowledges current limitations of automated systems in writing assessment and the criticism, including objections from NCTE, that machine scoring is unable to measure features such as meaningfulness of content or rhetorical effectiveness. Deane writes, "It is clear that e-rater (like most state-of-the-art AES engines) directly measures text quality, not writing skill" (p. 16). The author argues that potential uses for AWE should focus on the ways that fluency in production of text, a largely cognitive function, relates to the broader range of social and cultural practices of effective writers. Although AWE systems have been primarily used in large-scale standardized testing systems, Deane suggests that they also can play a role in those aspects of writing instruction focused on technical text production and in concert with other assessment systems, such as portfolios. The article also argues that AWE systems can help to identify and intervene with students for whom basic text production presents significant challenges.

KEYWORDS: machine-scoring, construct-validity, argumentation, rhetorical, production, cognitive-processing, research-agenda, sociocognitive, e-rater, learning-theory, ETS

Vojak, Colleen; Sonia Kline; Bill Cope; Sarah McCarthey; Mary Kalantzis

New spaces and old places: An analysis of writing assessment software

Computers and Composition 28.2 (2011), 97-111

The authors use sociocultural theories of learning to examine whether Automated Writing Evaluation (AWE) systems harness the potential of new technologies to promote broader conceptions of writing as a socially constructed and meaning-making activity. After a review of 17 systems, which included product demonstrations when available, the authors assert that existing programs primarily reinforce "old practices"—that is, narrow views of formal correctness in writing that align most easily with large-scale assessments. The authors' work is guided by questions about how well AWE systems fit with understandings of writing as a socially situated activity, writing as functionally and formally diverse, and writing as a meaning-making activity conveyed in multiple modalities. The authors find that while AWE systems include appealing features, such as swift feedback and plagiarism detection, the systems overall fail to reflect social and cultural constructs of writing. Specific findings include that the systems offer limited opportunity for pre-writing processes or collaboration, provide vague feedback that could confuse novice writers, and value formulaic conventions and length over true invention. The authors conclude that the primary concern is not the technology itself, but, rather, the restricted view of writing that underlies the systems. The authors conclude that emerging technology could help students develop as writers, but only if students are "provided with opportunity to explore the social contextuality, diversity, and multimodality of what it means to write in a digital age" (p. 109).

KEYWORDS: computer-analysis, Criterion, MY Access!, PEG, MyCompLab, Calibrated Peer Review, correlation, data, software-analysis, formulaic, false-flag, MX, essay-length

Williamson, David M.; Xiaoming Xi; F. Jay Breyer

A framework for evaluation and use of automated scoring

Educational Measurement: Issues and Practice 31 (2012), 2-13

Williamson, Xi, and Breyer suggest a theoretical model for the evaluation and implementation of Automated Writing Evaluation (AWE) systems that could provide WPAs with relevant guiding questions when considering use or expansion of machine-scoring systems. While the article focuses on evaluation of AWE systems intended to serve as a second rater in high-stakes assessments, the authors suggest that their proposed evaluation criteria also could have application in the use of automated evaluation in classroom practice or for lower-stakes testing. The authors propose five areas for evaluating AWE systems. These include how well the capabilities of the proposed system match the desired assessment and how closely the system aligns with human scoring. The authors also suggest exploring less commonly considered factors, including how well a system aligns with independent measures such as student grades and the consequences of AWE decisions on test-takers, especially those in non-dominant populations. The authors use ETS's e-rater® system to

illustrate their proposed evaluation framework, but the article addresses general principles concerning the structure and adoption of AWE systems. It also raises additional questions about AWE, such as how test takers adapt their behaviors when they are aware that their written responses will be scored at least partially by machine.

KEYWORDS: machine-scoring, assessment, e-rater, large-scale, scoring, entrance-requirement, validity framework, criteria, test-taker, behavior

Section 2: Scoring

Burstein (2013) defines Automated Writing Evaluation (AWE) as a technology that provides scores (assigning a number to an essay) or evaluation (providing diagnostic feedback) for written products. This definition is especially useful when combined with the realization that AWE systems do not read writing samples. Whether designed to score short-answer responses or longer writing samples, AWE computer programs do not read a writing sample as would a human. Rather, the systems are designed to predict a human rater's score within given boundary domains (Bennett, 2011). As such, for a given AWE system to predict human scores there must be resonance between the constructed-response task (Bennett, 1993) that elicits the semantic, syntactic, and lexical variety from the writer and the computational technique (Landauer, McNamara, Dennis, & Kintsch, 2007) used to capture such variety. At the present time, no AWE system claims to capture aspects of the writing construct associated with competencies such as audience awareness. As such, scoring and evaluation of aspects of the writing construct based on certain cognitive, intrapersonal, and interpersonal domains (National Research Council, 2012) must be scored by human readers.

Bennett, Randy Elliot

Automated scoring of constructed-response literacy and mathematics items

Washington, DC: Arabella Philanthropic Investment Advisors (June 2011)
http://www.ets.org/s/k12/pdf/k12_commonassess_automated_scoring_math.pdf

Complex constructed-response tasks designed to measure developed ability in both literacy and mathematics are amenable to Automated Writing Evaluation (AWE). However, scoring is only one component of an AWE system. Attention to the interplay of other system components is equally important. These components include construct definition, student interface, student tutorial, item development, scoring program design, and result reporting. Attention to scoring as part of an integrated process of validation provides a framework for responsible use. Users should: 1) view the use of AWE as a series of interrelated components; 2) encourage vendors to base scoring approaches on construct understanding; 3) call for studies that strengthen understanding of human scoring, especially the bases on which humans assign scores, as means of improving automated scoring; 4) stipulate vendor disclosure of scoring approaches; 5) require a broad base of validity evidence before adopting an AWE system; and, 6) unless validity evidence justifies exclusive use of automation, include human raters as part of the process when using AWE.

KEYWORDS: machine-scoring, AWE (automated writing assessment), consequence, scoring, guidelines, recommendations, validity framework

Burstein, Jill

Automated essay evaluation and scoring

In Chapelle, Carol A. (Ed.), *The encyclopedia of applied linguistics*; West Sussex, England: Wiley-Blackwell (2013), 309-315

From the earliest systems in the mid-1960s (Page, 1966) to the present, considerable technological advances have been made in Automated Writing Evaluation (AWE) systems. Burstein argues that two approaches have proven most fruitful: Natural Language Processing (NLP), a computational methodology that extracts linguistic features (semantic, syntactic, and lexical variety); and Latent Semantic Analysis (LSA), a corpus-based statistical modeling technique that uses large corpora to predict word use in a given subject domain. While NLP is based on a defined construct model of features in student writing, LSA is based on word use in reference documents such as textbooks. In practice, distinctions between the two approaches can be seen in e-rater® (developed by the Educational Testing Service) and its use of NLP and Write to Learn™ (developed by Pearson) and its use of LSA. Because they can rate and diagnose linguistic features of a written product, AWE systems can provide useful support for low- and high-stakes environments for teaching and assessing writing. Because of their design, the most sophisticated systems—those designed with a transparent modeling approach designed and improved through reported research—provide information on writing features used by humans in assessment settings. Additionally, AWE systems can reduce reporting times and costs associated with human scoring. Research programs currently underway will allow a larger variety of constructed-response tasks to be developed that can be used across wider populations.

KEYWORDS: construct model assessment methods, corpus-based modeling methods, Natural Language Processing (NLP), Latent Semantic Analysis (LSA), machine-scoring

Herrington, Anne; Charles Moran

Writing to a machine is not writing at all

In Elliot, Norbert; Perelman, Les (Eds.), *Writing assessment in the twenty-first century: Essays in honor of Edward M. White*; New York, NY: Hampton Press (2012), 219-232

Focusing on the Criterion® Online Writing Evaluation Service, an instructional and assessment system developed by the Educational Testing Service (ETS), the authors present a case study of "stumping" designed to demonstrate limits of the system's scoring and evaluation capability. (For an example of a large-scale stumping study, see Powers, Burstein, Chodorow, Fowles, & Kukich, 2001). Writing their own essay in response to a topic available on Criterion, the authors provide an analysis demonstrating that the system does not align with their conceptions of score range and definitions of flagged errors. The case study

illustrates two constraints of automated writing assessment: the limits of Standard American English itself and drawbacks of decontextualized writing.

KEYWORDS: audience, consequence, ESL, EFL, evaluation, computer-feedback, Criterion, Educational Testing Service, cost, machine-scoring, holistic, false-flag, dialect, standard written English, data, validity framework

Leacock, Claudia; Martin Chodorow; Michael Gamon; Joel Tetreault

Automated grammatical error detection for language learners

San Rafael, CA: Morgan and Claypool (2010)

Control of language is an important part of large-scale testing. Because Automated Writing Evaluation (AWE) scores should include detection of grammatical error, the authors provide an overview of Natural Language Processing tools, such as part-of-speech taggers and parsers, to explain how statistical methods can be used to detect error. Scoring processes featured are those of inter-reader agreement used to determine an overall score such as the quadratic weighted kappa, a measure of reader consensus (Stemler, 2004; Williamson, Xi, & Bryer, 2012.) Attention is given to methods of evaluating error detection systems, with special attention to precision (how many of the system's predictions were correct) and recall (how many of the errors were accurately identified by the system). The volume identifies longitudinal studies that address the degree to which English Language Learners benefit from human and automated corrected feedback. While primarily intended for those interested in Computer-Assisted Language Learning (CALL), the volume will also be of interest to those interested in English as a Foreign Language (EFL) and English as a Second Language (ESL).

KEYWORDS: Automated grammar error detection, grammar-checker, Computer-Assisted Language Learning (CALL), evaluation, scoring, error detection, EFL, ESL, statistical techniques, validity framework, longitudinal, data

McCurry, Doug

Can machine scoring deal with broad and open writing tests as well as human readers?

Assessing Writing 15.2 (2010), 118-129

This article reports the results of a field trial in which Automated Writing Evaluation (AWE) "did not grade broad and open writing responses as reliably as human markers" (p. 188). The author posits that widespread claims about the reliability (understood in this study to be inter-reader agreement) of machine scoring are derived from the assessment of "narrow and convergent tasks, and that they depend on such tasks to produce results that roughly mimic human judgments" (p. 119). To provide warrant for such claims, the author first analyzes a National Assessment of Educational Progress (NAEP) report from the Technology-based Assessment Project (Sandene, Horkay, Bennett, Allen, Braswell, Kaplan, & Oranje, 2005). Results from that study showed that the AWE system (in this case, e-rater®) did not agree

with scores awarded by human raters and produced mean scores that were significantly higher than the mean scores awarded by human readers. Human scores, the NAEP study also found, correlated more highly with one another than with the AWE scores, and the human raters assigned the same score to papers with greater frequency than the AWE assigned the score. After citing "relatively specific and constrained" (p.121) writing tasks from the Graduate Record Examination (GRE) Analytic Writing section, the author proposes that the AWE and reader agreement issues on the NAEP study may be due to writing task type. The author then presents the results of a case study of the writing section of the Australian AST Scaling Test (AST), a measure designed to reflect classroom practice task design. Scoring of the AST yielded adjacent scores in approximately 80% of the scored writing samples. Using two unidentified AWE platforms that had been modified on the basis of 187 AST samples, the author then used the system to score an additional 63 writing samples. Results revealed that same scores were awarded by humans on 36.9% of the writing samples, while the two AWE systems recorded exact agreement at 22.2% and 14.5%. In terms of same or adjacent scores, humans awarded these scores 88.9% of the time, while the AWE systems awarded these scores to 82.5% and 82.3% of the examined samples. The author concludes that the two AWE systems "differ significantly from those of the human markers of the AST Writing Test" (p. 127).

KEYWORDS: Australian AST Scaling Test (AST), machine-scoring, human-machien, data, automated essay scoring, computer scoring of writing, Graduate Management Admissions Test (GMAT), interrater-reliability, inter-reader reliability, National Assessment of Educational Progress (NAEP), writing task design

Shermis, Mark D.; Ben Hammer

Contrasting state-of-the-art automated scoring of essays: Analysis

Vancouver, BC: National Council of Measurement in Education (2012) http://www.scribd.com/doc/91191010/Mark-d-Shermis-2012-contrasting-State-Of-The-Art-Automated-Scoring-of-Essays-Analysis.

Also reported in Shermis and Hammer (2013), this study was conducted during a period when the viability of using Automated Writing Evaluation (AWE) systems was under consideration for assessing student performance under the Common Core State Standards Initiative (National Governor's Association, 2013). The study compared the results generated by nine AWE engines on eight essay scoring prompts. The prompts were drawn from six states that annually administer high-stakes writing assessments. Student essays from each state were randomly divided into three sets: a training set (used for modeling the essay prompt responses and consisting of text and ratings from two human raters along with a final or resolved score); a second test set used for a blind test of the vendor-developed model (consisting of text responses only); and a validation set that was not employed in the study. The essays encompassed writing assessment items from three grade levels (7, 8, 10). The essays were evenly divided between prompts using source material and prompts that did not require the use of sources. The total sample size for training the AWE systems was 22,029. The number of essays scored following training was 4,343. The results demonstrate that the

AWE systems were within 0.10 of the resolved mean scores of the human raters. As well, AWE systems produced scores that were equal or greater than human scores. The essay closes with qualifications of the study: that human rating is not necessarily the best standard to use in judging AWE systems; that the construct measured by the systems may not align with other construct measures; that the systems may be manipulated by test-takers; and that fairness for diverse populations remains an area of concern. For a critique of this research study, see Les Perelman (2013), below.

KEYWORDS: machine-scoring, automated essay scoring, validation, Common Core State Standards Initiative, high-stakes assessment, human-machine, interrater-reliability, data, Race to the Top Consortia

Section 3: Use

Claims and counter-claims are routinely made about the benefits and liabilities of Automated Writing Evaluation (AWE) based on principles of validation (Kane, 2006, 2013). However, often these claims are based on studies that do not focus on the ways this kind of evaluation is actually used. The articles included here address how AWE is used in classrooms, specific systems and features that might shape instruction, and contexts in which AWE might or might not be the most effective measure of student writing performance.

Balfour, Stephen P.

Assessing writing in MOOCs: Automated essay scoring and calibrated peer review

Research and Practice in Assessment 8 (2013), 40-48

The rapid expansion of MOOCs—massive, open, online courses—has raised questions about how instructors can provide feedback on writing in classes that are able to enroll tens or even hundreds of thousands of students. It also has created a possible area for swift growth in the use of Automated Writing Evaluation (AWE) systems. In this article, Balfour examines the contrasting approaches adopted by the two largest MOOC organizations. EdX, the non-profit organization formed by MIT and Harvard, announced in April 2013 that it would adopt an automated system to provide machine-based feedback on student writing in MOOCs. Meanwhile, Stanford-based Coursera has signaled its skepticism of AWE and continues to use a system of calibrated peer review in which students provide feedback to each other and on their own work using a rubric developed by the course instructor. In a discussion based on a literature review of existing AWE and calibrated peer review systems, Balfour argues that while AWE can provide swift and consistent feedback on some technical aspects of writing, peer review programs have been shown to help students increase confidence in their own composing and improve general learning skills such as the ability to evaluate material. The article suggests that MOOC organizers consider a blended evaluation approach in which students use AWE to address mechanical issues in their writing and employ calibrated peer review as a way to consider broader issues of content and style. Noting that researchers working for system developers dominate much of the existing AWE literature, Balfour also

suggests that MOOCs could offer a new site for more independent research into the use of automated evaluation programs.

KEYWORDS: machine-scoring, assessment, large-scale, MOOC, online, Calibrated Peer Review, EdX, Coursera, blended evaluation approach

Burstein, Jill; Beata Beigman-Klenanov; Nitin Madnani; Adam Faulkner

In Shermis, Mark D.; Jill Burstein (Eds.), *Handbook of automated essay evaluation*; New York, NY: Routledge (2013), 281-297

An emerging area of research, sentiment analysis borrows from systems of analysis used in specific domains such as movie reviews and political discourse in an effort to create a system that can identify sentiment and polarity (positivity, negativity, and neutrality) in the sentences of student essay writing. In the view of the authors, sentiment or feeling and its polarity or range of expression are relevant to the evaluation of argumentation in writing. That is, expression of personal opinion, attributed opinion and positive or negative statements about the topic under discussion help to build an argument. The authors describe in detail the methods by which they created and evaluated subjectivity lexicons or lists of words that can be categorized as evoking positive, negative or neutral feelings. Words such as "politeness," "lovely," and "fairness" are positive, and words such as "attacked," "misgivings," and "insufficiently" are negative, while words like "say," "there," and "desk" are neutral. These lexicons can, in the view of the authors, help to identify aspects of sentiment in essays produced by students in writing tests, and this identification can enhance the construct of writing addressed by machine scoring.

KEYWORDS: argumentation, Natural Language Processing, sentiment analysis, construct of writing

Cope, Bill; Mary Kalantzis; Sarah McCarthey; Colleen Vojak; Sonia Kline

Technology-mediated writing assessments: Principles and processes

Computers and Composition 28.2 (2011), 79-96

This companion piece to the Vojak, Kline, Cope, McCarthey, and Kalantzis (2011) analysis of writing assessment software surveys the current state of writing assessment, noting that it is largely summative and does not support student learning. It then posits that technology has the potential to effect a shift toward a greater emphasis on formative assessment and to foster more effective assessment in multiple disciplines. The authors propose six transformations that would make it possible for writing to become central in formative assessment across the curriculum. These transformations are: 1) assessment should be situated in knowledge-making practices like those enacted in various disciplines, and it should balance reading with writing; 2) assessment should measure social cognition rather than emphasize rote memory; 3) assessment should measure metacognition, which is essential for effective use of today's textual and knowledge environments; 4) assessment should be conducted in spaces where

learners can represent their knowledge multimodally; 5) assessment should draw upon peer review as well as teacher monitoring to provide rapid formative assessment; 6) assessment should capitalize on the ubiquitous presence of computing capacity to create a great number of formative assessments with the goal of abolishing the functional difference between formative and summative assessment. In addition, this article argues for bringing together these six technology-mediated processes for assessing writing—natural language analytics, corpus comparison, in-text network-mediate feedback, rubric-based review and rating, semantic web processing, and survey psychometrics—in order to link formative and summative assessment.

KEYWORDS: assessment, evaluation, technology, guidelines, contextual, knowledge-making, metacognitive, multimodal, learning-theory, computer-analysis, summative, transformative, peer-evaluation, formative

Deane, Paul; Frank Williams; Vincent Weng; Catherine S. Trapani

Automated essay scoring in innovative assessments of writing from sources

Journal of Writing Assessment 6 (2013)
http://www.journalofwritingassessment.org/article.php?article=65

With the advent of the Common Core State Standards and their increased demand for writing from source material in K-12 instruction, questions arise about what role automated scoring systems can play in evaluating such tasks in large-scale assessments. The authors argue that Automated Writing Evaluation (AWE) systems can provide measures of student performance on technical aspects of such writing, but they also indicate that human raters are needed to evaluate more complex factors such as critical reasoning, strength of evidence, or accuracy of information. The article reports on pilot studies from 2009 (a convenience sample of 2,606 8th grade students) and 2011 (a convenience sample of 2,247 8th grade students) of an Educational Testing Service project that addressed source-based writing. In both pilots, eighth-grade students from across the country completed writing tasks within the CBAL (Cognitively-Based Assessments of, for, and as Learning) digital platform on a series of complex writing tasks: argumentation, literary analysis, policy recommendations, and design of an information pamphlet (Deane, Fowles, Baldwin, & Persky, 2011). The researchers found that it is possible to train an operational AWE model to score writing from sources. In addition, they also found that the pattern of results suggests that the e-rater model validly addresses some aspects of writing skill (such as the ability to produce well-structured text fluently without major errors in grammar, usage, mechanics, or style), but not others (such as the ability to justify a literary interpretation). Emphasizing the significance of a combined automated and human scoring process, the researchers found that strength of the association between the automatically-scored and the human-scored parts of the test could support a reporting model in which the selected-response and automatically-scored essay portions of the test were used to provide interim score reports for formative purposes, that could later be combined with human scores. The findings are timely because the consortia developing Common Core assessments have indicated their intention to incorporate automated scoring to evaluate student writing, including writing from sources. Here, the authors draw attention to

the complexity of the writing construct involved in such tasks and suggest multiple measures are necessary to offer the most complete assessment, and they propose a combination of automated and human score use.

KEYWORDS: machine-scoring, Common Core Standards, measurement, argumentation, reasoning, sources, G8, data, interrater-reliability, human-machine, plagiarism, keystroke-analysis, human-machine, CBAL (Cognitively-Based Assessments of, for, and as Learning)

El Ebyary, Khaled; Scott Windeatt

The impact of computer-based feedback on students' written work

International Journal of English Studies 10.2 (2010), 122-144

Ebyary and Windeatt add to the relatively small body of research on the usefulness of computer-based feedback on student writing with an investigation of the effects of feedback provided by Criterion® Online Writing Evaluation System on prospective English as a foreign language (EFL) teachers enrolled in a large class (over 800 students) where instructor feedback was limited. The initial survey and interviews showed that only 17% of the students felt positive about the feedback they were receiving from instructors. The 24 student volunteers selected to participate were given four topics to write about during an eight-week period and for each submitted an initial draft and a revision using the computer-generated feedback. After receiving computer-generated feedback, 88% of these students expressed positive attitudes about the feedback they had received and about the quality of their own writing. Although use of the Criterion system had no discernible effect on the prewriting of study participants, it did increase their reflection on their own writing. Criterion scoring of students' drafts and revisions showed that computer-generated feedback had a positive effect on writing quality. Human scoring of these pieces of writing supported this finding. However, closer scrutiny of the student writing showed that a number achieved higher scores through avoidance strategies, avoiding language that would cause them to make errors in the usage, mechanics, and style measured by Criterion. Furthermore, students in this study had no basis for evaluating computer-generated feedback since they had previously received little or no feedback from instructors.

KEYWORDS: computer-generated feedback, EFL, ESL, Criterion, avoidance strategies, data, gain, evaluation

Scharber, Cassandra; Sara Dexter; Eric Riedel

Students' experiences with an automated essay scorer

The Journal of Technology, Learning and Assessment 7.1 (2008), 4-44

The authors describe a study in which students in a teacher education program were asked to use an electronic system that presented them with classroom cases or scenarios to which they responded with written essays. Formative assessment took the form of automated responses

to first drafts of these essays. These responses, based on a rubric established for the essays, generated an estimated score or grade for the draft and provided an explanation of a "good" response. The features measured by the automated scorer included vocabulary, usage, specific phrases, and grammatical structure. Data gathered to measure students' responses to this electronic formative assessment included pre- and post-assessment surveys, logs of students' actions while completing the assignment, and interviews with selected students. The surveys, which asked questions about students' experiences with computer technology and their beliefs about the role of technology in education, were used, along with log data, to contribute to rich portraits of students who were selected for interviews. Comparison of computer-generated formative assessment with instructors' summative evaluations showed that the Automated Writing Evaluation (AWE) system undervalued student writing in comparison to the instructor, and this finding may help account for the fact that students made less use of the AWE system's formative assessment on the second of the two writing tasks included in this study. The four students who were interviewed had mixed responses to the AWE system, but all expressed strong emotions about this system of formative evaluation, which highlights the importance of considering students' affective responses to AWE. Furthermore, those interested in improving the machine-based formative assessment of student writing should be guided by research on responding to student writing.

Keywords: machine-scoring, computer-analysis, feedback, pedagogy, pre-service, student-opinion, interview, corruptibility, formative, tricking

Warschauer, Mark; Douglas Grimes

Automated writing assessment in the classroom

Pedagogies: An International Journal 3.1 (2008), 22-36

In an effort to determine the effects of Automated Writing Evaluation (AWE) use on students and teachers in middle school and high school, the authors used interviews, surveys and classroom observations to study how the Criterion® Online Writing Evaluation System and the MY Access!® system were actually used in schools and their effects on teacher behavior and student writing. Although teachers claimed to value this software, they were not frequently used. This was in part because teachers felt pressure to prepare students for state exams. It also was because the programs could only provide feedback on essays written to the specific prompts that are part of the program, and teachers preferred more meaningful and authentic writing assignments. Teachers did report that these programs saved grading time, but they did not change teachers' habits and beliefs in significant ways. Software developers often claim that using AWE motivates students to revise more frequently, but this study found that most students submitted their papers for scores only once. Students recognized that making minor changes was the easiest way to improve their machine-produced scores, and nearly all revisions addressed spelling, word choice or grammar rather than substantial issues like content or organization.

KEYWORDS: middle-school, high-school, teacher behavior, student behavior, machine-scoring, essay-analysis, pedagogy, motivation, Criterion, My Access, grading time, machine-produced scores, student-opinion, teacher-opinion, data

Section 4: Diversity

Few studies of Automated Writing Evaluation (AWE) have disaggregated findings by sub-group to determine how the use of these systems affects diverse student populations. However, as Elliot, Deess, Rudniy, & Joshi (2012) suggest, the possibility of disparate impact has legal as well as ethical implications for programs and institutions considering the adoption of AWE products for admissions, assessment, or instruction. This section highlights the handful of studies that have examined the impact of AWE on diverse student populations, investigating AWE-related outcomes by gender, ethnicity, nationality, and primary language/English variety.

Bridgeman, Brent; Catherine Trapani; Attali Yigal

Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country

Applied Measurement in Education 25.1 (2012), 27-40

In this article, researchers examine differences by gender, ethnicity, and nationality between human and Automated Writing Evaluation (AWE) scores on two high-stakes timed essay tests that use ETS's e-rater® software: the Test of English as a Foreign Language (TOEFL) iBT and the Graduate Record Exam (GRE). These studies draw on the largest dataset of any study of AWE and diverse populations: 132,347 TOEFL essays and a random sample of 3,000 GRE "Issue" and "Argument" essays. In the TOEFL study, the authors found that, on average, e-rater scored writing by Chinese and Korean speakers more highly than did human raters, but gave lower scores to writing by Arabic, Hindi, and Spanish speakers. The authors hypothesize that these scoring differentials are attributable to stylistic differences that human readers often accommodate but e-rater does not; some of these differences may be cultural rather than linguistic. The TOEFL study showed no significant differences in human versus e-rater scores by gender. On the GRE, human scores for "Issue" essays by African American and Native American men were slightly higher than e-rater scores, and African American men and women both received higher human scores on the "Argument" essay. Among international GRE-takers, students from mainland China received higher scores from e-rater than from human readers, although the same difference did not hold for Mandarin-speakers in Taiwan, suggesting cultural rather than linguistic causes for the disparity. The authors hypothesize that e-rater might assign greater value to length and less value to certain grammatical features than many human readers, and that human readers might be more sensitive to whether an otherwise well-constructed essay is off-topic. The authors conclude that these sub-group differences suggest ways that AWE software can be refined, but note that any disparate impact from the current iteration of e-rater is muted by the GRE and TOEFL scoring model, in which major discrepancies between e-rater and human scores on individual essays are settled by a second human reader.

KEYWORDS: machine-scoring, human-machine, data, Graduate Record Examination, TOEFL, e-rater, Educational Testing Service, NS-NNS, international, ESL, EFL, bias, off-topic, arrangement, criteria, correlation, differential impact, ethnicity, gender-difference, race, African-Am, Native-Am

Elliot, Norbert; Perry Deess; Alex Rudniy; Kamal Joshi

Placement of students into first-year writing courses

Research in the Teaching of English 46.3 (2012), 285-313

This article presents the findings of a local validation study of several writing placement tests at an urban, public, research university dedicated to science and technology education. Examining concurrent and predictive evidence regarding writing placement tests used at the university over a twelve-year period (1998-2010), this study compares the relative ability of ACCUPLACER® (which evaluates student writing samples using WritePlacer *Plus*® software) and the human-scored SAT Writing Section to predict students' writing course grades and end-of-semester portfolio scores. The extrapolation portion of the study, which looked for differential impact of these exams on diverse student populations, showed that neither ACCUPLACER nor the SAT Writing Section were statistically significant predictors of Black or Hispanic students' first semester composition course grades. While the SAT Writing Section was predictive of second-semester course grades for all ethnic and gender groups, ACCUPLACER did not predict the second-semester grades of Asian, Black, Hispanic, or female students; however, both exams were predictive of White and male students' grades and portfolio scores across both semesters. The authors suggest that ACCUPLACER might not capture the complexities of the discourse features in essays by students from some groups. Because of the vendor's unwillingness to provide any kind of analysis of differential impact on diverse student populations, the authors concluded that continued used of ACCUPLACER could not be legally defended under Title VI and VII of the Civil Rights Act. Based on this study, the university decided that the SAT Writing Section was a better placement test for their students and discontinued use of ACCUPLACER.

KEYWORDS: placement, predictive, data, large-scale, ACCUPLACER, SAT-testing, validity, machine-scoring, minority, gender-difference, Hispanic, African-Am, data, local, idiographic, nomothetic, legality, bias, race, ethnicity, local

Herrington, Anne; Sarah Stanley

Criterion: Promoting the standard

In Inoue, Asao B.; Mya Poe (Eds.), *Race and writing assessment*; New York, NY: Peter Lang (2012), 47-61

In this chapter, Herrington and Stanley examine the language ideologies underpinning Criterion® Online Writing Evaluation System, an e-rater®-based Automated Writing

Evaluation program designed for instructional use. Drawing on the principles of NCTE/CCCC's policy statements regarding language diversity, as well as the growing body of scholarship challenging myths of linguistic homogeneity in and beyond the composition classroom, the authors interrogate Criterion's marketing materials and evaluation criteria and, as a case study, examine how Criterion evaluates an essay making use of African American discursive resources. The authors argue that although the images of students in Criterion's marketing materials are racially diverse, the corpus of human-scored essays from which e-rater derives its linguistic patterning is evaluated according to a single dialect standard: Edited American English. Furthermore, e-rater does not evaluate the full range of genres and rhetorical approaches or the kinds of critical thinking that students are asked to produce in classroom contexts. Neither the holistic nor the trait scoring features of Criterion acknowledge the existence of multiple rhetorical and linguistic options, and the feedback Criterion provides is often unhelpful or inaccurate. Likewise, the software often fails to recognize effective rhetorical and stylistic decisions, particularly when those decisions derive from alternative—often, raced and classed—discursive traditions. Herrington and Stanley find that Criterion promotes an arhetorical, error-focused construct of writing that recognizes only one dialect. While they acknowledge that some writing instructors share these ideologies, the authors assert that Criterion hinders the field's efforts to be self-reflective about who is privileged and who is marginalized by the promotion of a single language "standard." Ultimately, they argue, Criterion fails to recognize the full range of discursive options available to students in a linguistically diverse society.

KEYWORDS: machine-scoring, Educational Testing Service, Criterion, standards, race, African-Am, language diversity, e-rater, bias, ethnicity, Standard American English

James, Cindy L.

Electronic scoring of essays: Does topic matter?

Assessing Writing 13.3 (2008), 80-92

This article presents findings from a study of the impact of essay prompts, a phenomenon called *topic effect*, on ACCUPLACER® WritePlacer *Plus*® scores, with a particular focus on prompt-related differences by gender and native language. Using both the WritePlacer *Plus* Automated Writing Evaluation software and trained human readers, James evaluated 77 sample essays written in response to three different randomly assigned ACCUPLACER prompts at a public university. She found that, overall, there was no topic effect across the three prompts. There were statistically significant differences in WritePlacer *Plus* scores by topic among women, who scored significantly higher on one of the three essay prompts, although there were no differences by topic among men, and there were no significant differences in WritePlacer *Plus* scores by topic between men and women. Among non-native speakers of English, there were no significant differences in WritePlacer *Plus* scores by topic, although there were significant differences in overall score between native and non-native speakers, with native speakers receiving higher scores. Human scorers' evaluations aligned closely with WritePlacer *Plus* scores and did not reveal any topic effect. James concludes that, except among women writers in the study, there is little evidence of a

WritePlacer *Plus* topic effect, at least within this small, single-institution sample. However, she also notes that only three of the eleven available ACCUPLACER essay topics were evaluated and warns that larger sample size, different institution type, and/or other student populations might yield different results.

KEYWORDS: machine-scoring, prompt, bias, gender-difference, ESL, local, data, Educational Testing Service, Criterion, standards, race

Weigle, Sarah Cushing

English language learners and automated scoring of essays: Critical considerations

Assessing Writing 18.1 (2013), 85-99

This article uses the TOEFL iBT, powered by e-rater®, as a case study of the validity issues surrounding the use of Automated Writing Evaluation (AWE) software for English language learners. Weigle asserts that, although AWE software is generally designed to evaluate native English writing, the largest and fastest-growing market for AWE may be non-native English speakers learning English in non-English-speaking countries. This suggests the need to validate AWE systems specifically for non-native populations. Depending on the context, second-language writing assessment may prioritize either learning to write or using writing to teach content, and in many cases that content includes second language proficiency. To date, however, AWE systems have done little to distinguish between these kinds of assessment. Based on her own study of the TOEFL, Weigle concludes the following: 1) correlations between the overall e-rater scores and human scores were as high or higher than correlations between two human ratings; 2) e-rater was somewhat more consistent across prompts than human raters; 3) scores produced by e-rater and human scores were moderately correlated with other measures of writing ability (e.g. course grades and instructor feedback); and 4) the feature scores used to generate total scores differed across the two examined prompts. This last finding suggests the need for further study in the following areas: evaluation (the extent to which the computer-generated scores of ELL students can be taken as accurate representations of performance); generalization (the extent to which these scores provide appropriate estimates of student scores obtained from other, similar performances); explanation (the extent to which scores can be attributed to the defined construct); extrapolation (the meaningfulness with which scores indicate performance to the target domain); and utilization (the usefulness of scores for decision-making). Weigle argues there may be valuable applications for AWE when working with second-language English writers. However, she warns, AWE also has the potential to further marginalize these students, and all decisions about its use with second-language populations should be made with an awareness of the software's limitations.

KEYWORDS: machine-scoring, ESL, EFL, international, needs-analysis, TOEFL, e-rater, validity, data, native-nonnative

Section 5: Consequence

The increasing use of Automated Writing Evaluation (AWE) software has impacts on student placement, high school and college curricula, and the further development of large-scale writing assessments. Considering Messick's (1989) emphasis on the importance of evidential and consequential validity, robust discussions about the consequences of AWE are critical. This section of the annotated bibliography highlights current understandings of these consequences as documented in the research literature. Because of their potentially wide-ranging impacts, the development of AWE systems and the consequences of implementing AWE in local contexts are of concern to parents, teachers, students, administrators, and policy makers.

Cheville, Julie

Automated scoring technologies and the rising influence of error

The English Journal 93.4 (2004), 47-52

Cheville argues that the increasing use of Automated Writing Evaluation (AWE) software in k-12 education and the rising influence of large-scale writing assessments will make teachers become "data managers" with "less time and authority to decide what their students know and need" (p. 49). Integrating systems such as the Criterion® Online Writing Evaluation System or other AWE software into teacher education programs jeopardizes approaches to teaching that view language and writing as transactional rather than as formulaic. The article recounts Cheville's visit to the Educational Testing Service (ETS) campus to discuss a potential partnership between ETS and the School of Education at Rutgers. The partnership would have enabled the Rutgers faculty to incorporate ETS's Criterion into their teacher education program. Cheville opposes this partnership because integrating Criterion into the teacher education program would not only enervate education students' critical understanding of the politics of writing assessment but also "undermine the language and learning of their future students" (p. 48). Most of the article documents three main consequences that are likely to result from the increasing influence of AWE systems: 1) instruction in formulaic writing will become increasingly significant, while context-specific meaning will become less so; 2) increasing instructional time spent addressing how to avoid trivial errors (i.e., errors that are noted by Criterion but not necessarily by human readers); and 3) curricula will have less space where writing teachers can respond to students' work in complex, transactional ways. In addition, the article presents a critique of one specific feature within Criterion—the portfolio. Cheville argues that the way "portfolios" have been incorporated into Criterion is unlikely "to facilitate reflective transfer" (p. 49). That is, the mechanized feedback built upon a collection of a student's writing samples is not likely to lead to the forms of reflection advocated by proponents of portfolios.

KEYWORDS: machine-scoring, ETS, Criterion, pilot, social, constructivist, assessment, high-stakes, error, power, learning-theory, K-12, teacher-evaluation, translational, Rutgers University, portfolio

Haudek, Kevin C.; Jennifer J. Kaplan; Jennifer Knight; Tammy Long; John E. Merrill; Alan Munn; Ross H. Nehm; Michelle Smith; Mark Urban-Lurain

Harnessing technology to improve formative assessment of student conceptions in STEM: Forging a national network

CBE Life Science Educational 10.2 (2011), 149-155

This meeting report presents a summary of work being done on how software systems may be used to evaluate students' "constructed-responses" (i.e., writing students do in response to a tightly delimited and usually timed prompt, as in examination questions) about discipline-specific concepts in six different curricular areas in Science, Technology, Engineering, and Mathematics (STEM). The report includes summaries of work on student learning in cellular metabolism, evolution and natural selection, genetics, introductory biology, geosciences, and statistics. These National Science Foundation (NSF)-funded projects began with the recognition that students' written responses may better reveal students' understanding (or misunderstanding) of key science concepts than multiple-choice assessments. One of the challenges for science educators has been that student constructed-responses are "time- and resource-intensive to evaluate" (p. 149). The working group has turned to text analysis software to see if there are means available to evaluate constructed-responses accurately and efficiently. They investigated two different software tools: SPSS Text Analytics for Surveys (STAS) and the Summarization Integrated Development Environment (SIDE). STAS is based on qualitative research software; it "extracts lexical tokens that can be used to create categories and rules using an analysis model similar to open coding in grounded theory" (p. 153). SIDE uses a machine learning scoring model; it "takes a set of human-scored responses and 'discovers' word patterns that account for human-generated scores" (p. 151). The group's research foci included: 1) questioning whether constructed-response items are always more effective in uncovering student thinking than multiple-choice items; 2) the generalizability of discipline-specific lexical analysis protocols; 3) strengths and weaknesses of qualitative research software compared with machine learning software; 4) relationship of software scoring to expert human scoring; 5) relationship of text analysis and scoring rubrics; 6) possibilities of using linguistics to enhance discipline-specific work on lexical analysis; and 7) establishing a large-scale data collection plan across the participating universities. Comparing STAS and SIDE, the authors note that SIDE (the machine learning software) offers a major time advantage, while STAS (the qualitative research-based software) "offers the advantage of discovering novel ideas by exploring students' use of language" (p. 153) rather than relying on models predicated on expert readers' scores. The meeting report concludes by sketching directions for future research, including the hope "to develop a web portal where users could upload their own sets of student responses and receive formative feedback in near real-time" (p. 154). The consequences of this work could be an increased use of Automated Writing Evaluation systems to encourage and evaluate WID within STEM fields.

KEYWORDS: machine-scoring, evaluation , computer-analysis, formative, student-conception, science-writing, pedagogy, ideas, evaluation, data

Klobucar, Andrew; Paul Deane; Norbert Elliot; Chaitanya Ramineni; Perry Deess; Alex Rudniy

Automated essay scoring and the search for valid writing assessment

In Charles Bazerman; Charles Dean; Jessica Early; Karen Lunsford; Suzie Null; Paul Rogers; Amanda Stansell (Eds.), *International advances in writing research: Cultures, places, measures*; Fort Collins, CO: WAC Clearinghouse; Anderson, SC: Parlor Press (2012), 103-120

Klobucar et al. argue that the ever-expanding assortment of digital writing technologies makes it likely that automated assessment technologies, including Automated Writing Evaluation (AWE) systems, will play an increasing role in both writing instruction and assessment. They urge WPAs to consider AWE as one tool among many that could be used in building effective postsecondary writing assessment systems. They argue for the importance of local validation activities for all writing assessment tools, including AWE systems, used at a university. Moving from a call for more local validation studies to their own descriptive case work, Klobucar et al. present a study from NJIT that analyzes how Educational Testing Service's Criterion® Online Writing Evaluation System performed when evaluating the writing of first-year students. Scores from multiple writing assessment instruments (i.e., SAT-W scores, scores on the two AWE-scored essays in Criterion, course grades, and holistic traditional portfolio scores) were analyzed. Klobucar et al. find that "Criterion can be used as an early warning system for instructors and their students" (p. 110); however, their study also indicates that having multiple measures and wide construct coverage is vital for fair and accurate assessments. Different writing assessment tools tap into different domains, with each one only partially capturing information about overall student performance. In the local context of first-year writing at NJIT, the constructs measured by Criterion appear to be necessary, but not sufficient to achieve success in the first-year writing course. Given their carefully defined context and purpose for using Criterion, Klobucar et al. conclude that this particular AWE software used for the purpose of identifying students in need of instructional support is "relatively strong" (p. 113).

KEYWORDS: machine-scoring, validity, evaluation, ETS, New Jersey Institution of Technology, Criterion, SAT-testing, pedagogy, error, MX, multiple-measures

Perelman, Les C.

Critique of Mark D. Shermis & Ben Hammer, "Contrasting State-of-the-Art Automated Scoring of Essays: Analysis"

Journal of Writing Assessment 6 (2013)
http://www.journalofwritingassessment.org/article.php?article=69

Perelman responds to Shermis and Hammer's (2012, 2013) comparative study of Automated Writing Evaluation (AWE) systems (see above). Perelman is critical of the study's original status and release as a white paper rather than as a peer-reviewed journal article. In his *Journal of Writing Assessment* study, Perelman refutes the claim that AWE systems are as accurate when scoring student writing as human readers. He argues: a) that Shermis and

Hammer do not present a clearly articulated construct of writing; b) that their methodology is flawed; and c) that their conclusions are impressionistic rather than based on statistical tests. Perelman's major contribution to AWE is his call for disaggregation of large data samples. Although the Shermis and Hammer study claimed that it was exploring how well machines could score extended-response writing associated with essays, only three of the eight datasets consisted of what is commonly defined as extended-response writing (i.e., had average word lengths of over 360 words). The mean number of words for the other five datasets ranged from 98.70 to 173.43 words. Further, data disaggregation revealed that human scorers performed at levels equal to or better than the AWE systems for most of the datasets. In cases where datasets are large and assumptions are made regarding claims that will have policy impact, Perelman argues that researchers should publicly post their data for analysis by other researchers so that the original findings may be confirmed, qualified, or refuted.

KEYWORDS: Mark D. Shermis, Ben Hammer, Contrasting State-of-the-Art Automated Scoring of Essays, critique, statistical-analysis, human-machine, interrater-reliability, machine-scoring, essay-length, validity, reliability, placement, ACCUPLACER, ETS, Criterion, e-rater

Ramineni, Chaitanya

Validating automated essay scoring for online writing placement

Assessing Writing 18.1 (2013), 40-61

Ramineni argues that a version of the Criterion® Online Writing Evaluation System software that uses customized scoring models can facilitate placement of students into writing courses. Her study reviews the history of writing assessment and the advocacy work for direct assessments of students' writing skills rather than indirect measures. She discusses how Automated Writing Evaluation (AWE) systems build both standardized, pre-existing models on general samples and how AWE systems may be tailored to a particular student population. Her study compares how customized, prompt-specific Criterion models compare with standardized, pre-existing models, considers the predictive validity of e-rater® scores, and examines the performance of e-rater across different testing conditions. She finds that the customized, prompt-specific Criterion models out-perform standardized, pre-existing scoring systems; that the predictive performance of e-rater had a positive correlation with the students' writing course grades, portfolio scores, and cumulative GPAs; and that this study's comparison of the impact of proctored and unproctored uses of Criterion was not conclusive. Ramineni sees the consequences of using Criterion reaching beyond merely facilitating placement into advanced, first-year, or basic writing courses; she advocates for the use of AWE systems as tools to offer WPAs additional information about student performance as well as potentially providing additional feedback directly to students.

KEYWORDS: machine-scoring, Criterion, placement, program, prompt-specific, predictive, validity, data

References

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), Construction vs. choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 1-27). Hillsdale, NJ: Erlbaum.

Bennett, R. E. (June, 2011). Automated scoring of constructed-response literacy and mathematics items. Washington, DC: Arabella Philanthropic Investment Advisors. Retrieved from http://www.ets.org/s/k12/pdf/k12_commonassess_automated_scoring_math.pdf

Burstein, J. (2013). Automated essay evaluation and scoring. In Chapelle, C. A. (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 309-315). West Sussex, UK: Wiley-Blackwell.

Council of Writing Program Administrators, National Council of Teachers of English, and National Writing Project (2011). *Framework for Success in Postsecondary Writing.* Retrieved from http://wpacouncil.org/framework.

Deane, P., Fowles, M., Baldwin, D., & Persky, H. (2011). *The CBAL summative writing assessment: A draft eighth-grade design* (Research Memorandum 11-01). Princeton, NJ: Educational Testing Service.

Elliot, N., Deess, P., Rudniy, A., & Joshi, K. (2012). Placement of students into first-year writing courses. *Research in the Teaching of English*, *46*, 285–313.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement.* (4th ed., pp. 17-64). Westport: CT: American Council on Higher Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum.

Messick, S. (1989) Validity. In R. L. Linn (Ed.) , *Educational measurement* (3rd ed., pp. 13-103). Washington , DC : American Council on Education & National Council on Measurement in Education.

National Council of Teachers of English Tasks force on Writing Assessment (2013). Machine scoring fails the test. Retrieved from http://www.ncte.org/positions/statements/machine_scoring

National Governors Association (2013). National Governors Association (2013). *Common core state standards initiative*. Washington, DC: National Governors Association. Retrieved from http://www.corestandards.org

National Research Council (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: National Academies Press.

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

Powers, D. E., Burstein, J., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). Stumping *e-rater*®: Challenging the validity of automated essay scoring (GRE No. 98-08bP, ETS RR-01-03). Princeton, NJ: ETS.

Sandene, B., Horkay, N., Bennett, R. E., Allen, N., Braswell, J., Kaplan, B., et al. (Eds.) (2005). *Online Assessment in mathematics and writing: Reports from the NAEP Technology-Based Assessment Project*. Research and Development Series, National Assessment of Educational Progress. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Retrieved from http://www.ecs.org/html/Document.asp?chouseid=6337

Shermis, M. D., & Hammer, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), Handbook of automated essay evaluation (pp. 313-346). New York, NY: Routledge.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4). Retrieved from http://PAREonline.net/getvn.asp?v=9&n=4

Vojak, C., Kline, S., Cope, B., McCarthey, S., & Kalantzis, M. (2011), New spaces and old places: An analysis of writing assessment software. *Computers and Composition*, *28*, 97-11.

Williamson, D. D., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practices*, *31*, 2–13.