Title
Automated Writing Evaluation

Your Name
Elena Cotos

Affiliation
Iowa State University

Email Address
ecotos@iastate.edu

Abstract
Automated Writing Evaluation (AWE) comprises a suite of web-based applications for computer-assisted assessment and learning. This historically controversial technology, solidly grounded in psychometric research, imposes the need for comprehensive inquiry into its context-specific utilizations in order to exploit its advantages appropriately and to devise effective classroom techniques for fostering writing development.

Main Text

Framing the Issue:
The term Automated Writing Evaluation (AWE) denotes a form of intelligent instructional technology as well as a multi-disciplinary field, which intertwines research and development informed by applied linguistics, computer science, educational measurement, writing studies, and psychometrics. In its fifty-year history, the field has germinated from inventing stand-alone automated scoring engines to designing complex interactive systems that support the writing process. The core function underlying AWE is its computational ability to evaluate written discourse. Relying on a combination of artificial intelligence and statistical models trained to extract a variety of linguistic features, AWE systems analyze written texts and produce output that translates to holistic scores assessing writing quality and to feedback aimed at facilitating improvement.

AWE emerged in the 1960s, when the *Project Essay Grade (PEG)* program was created to address a major challenge in writing instruction – the time-intensive, daunting task of grading writing assignments, which was also viewed as an impediment to providing sufficient practice opportunities essential for writing development. In the 1990s, advances in computer technologies provided a fruitful platform for computational analysis of naturally occurring language and, therefore, for electronic processing of constructed written responses. The Internet offered massive accessibility opportunities. Together, these factors contributed to the development and employment of engines such as *e-rater*, *Intelligent Essay Assessor (IEA)*, *Intellimetric*, *Constructed Response Automated Scoring Engine (CRASE)*, *AutoScore*, *Bookette*, etc. Their original intended use was to score essays by emulating human scoring behavior. Complementing human ratings, these systems found direct implementation in large-scale tests to provide immediate, individualized, and consistent assessment. For example, *e-rater* was integrated for operational scoring of writing tasks in the Internet-based Test of English as a Foreign Language (TOEFL iBT®), *IEA* in the Pearson Test of English (PTE), and *Intellimetric* in the Graduate Management Admissions Test (GMAT).

Automated scoring extended to the generation of feedback; *Criterion*, *WriteToLearn*, *MyAccess!,* and *Writing Power* are augmentations of *e-rater*, *Intelligent Essay Assessor*, *Intellimetric*, and *CRASE*, respectively. This evolution is reflected in several iterations of the name – from automated essay scoring

(AES) to automated essay evaluation (AEE) and to the more encompassing automated writing evaluation (AWE) term, with *writing* potentially representing any genre, and *evaluation* indicating extrapolation of computational outputs to uses other than scoring. Recently, AWE has crossed the essay boundary, stepping into the realm of research genres. As a rule, AWE programs analyze student drafts and return feedback related to various writing traits (e.g., language errors, usage and mechanics, syntactic complexity, variation in sentence type, style, development of ideas, conceptual content, topic relevance, discourse structure, and rhetorical functions).

Although analysis engines are used for both scoring and feedback, these are not interchangeable concepts. AES systems, generally used to assess writing performance for important decision-making, are not intended for direct deployment in the classroom. It is their feedback-generating derivatives that are designed for classroom use. Therefore, making a clear distinction between systems built for summative assessment purposes and those designed for formative assessment is of utmost importance when it comes to implementation. Considerate use of AWE capabilities can play a significant role for teaching and learning, while misuse can lead to inappropriate interpretations and generalizations that may result in undesirable implications.

Making the Case

Since the inception of AWE, heavy doubts have been casted on these systems, and the skepticism magnified with the dual testing-instruction application of scoring engines. The developer-researchers, while making earnest efforts to enhance writing praxis with innovative technology, have been blamed for being caught in the world of 'everything is measurable,' not accounting for the theoretical essentials underpinning writing pedagogy. The in-situ writing community rejects any application that is based on automated processing, criticizing the very idea of having computers perform tasks that require human intelligence and refusing to acknowledge the potential usefulness of AWE. Part of the problem is that writing practitioners and educational measurement professionals have not yet developed a transparent definition and a mutual interpretation that would bridge the two epistemologies to inform computational operationalization of the writing construct.

Automated scoring scholars ground their systems in cognitive information-processing, treating writing as a construct that contains quantifiable features, which, when measured in the aggregate, can help make relatively accurate predictions of holistic human evaluation. These features, treated as indicative of specific aspects of writing performance, are modeled on human ratings guided by scoring rubric criteria for specific writing tasks. Therefore, most empirical inquiries aimed to demonstrate the potential of automated scoring have been psychometric in nature, focusing on agreement between automated and human scores and providing evidence of relatively high reliabilities between AES and human raters or other comparable measures (Shermis & Burstein, 2003).

Writing practitioners are not persuaded by this kind of evidence; on the contrary, they contend against such deterministic evaluation of writing. The underlying argument is that the writing construct can by no means be dissected into formulas and evaluated computationally because writing is a deeply human and creative activity. From their perspective, the writing construct includes "the rhetorical ability to integrate an understanding of audience, context, and purpose when both writing and reading texts; the ability to think and obtain information critically; the ability to effectively employ multiple writing strategies; the ability to learn and use the conventions appropriate to a specific genre of writing; and the ability to write in various and evolving media" (Perelman, 2012, p. 129). Since computers cannot understand meaning, replicate cognitive processes involved in human evaluation of texts, and incorporate the communicative dimensions of writing, automated scoring is viewed as undermining teachers' ideological beliefs and even wanted expelled from writing classrooms (Cheville, 2004).

These theoretical and research considerations have important implications for writing practice, which both computational and humanistic perspectives aim to enhance. Unlike the former, which values

empirical methods for linguistic description, explanation, and modeling, the axiology of writing studies is on influencing writers' attitudes and behaviors, with inquiry methods centering on interpretations that draw from interaction between researchers and participants. The irreconcilable can become reconcilable if AWE evaluation research connects interdisciplinary interests by extending the points of inquiry to the ecology of the targeted contexts, especially ESL/EFL. To overcome barriers inevasible for most educational innovations, it is imperative to acquire a comprehensive understanding of the learning potential of AWE and the effects it may exert on teaching and learning.

Pedagogical Implications

Drastic views regarding AWE contributed to the emergence of an evaluative user-centric strand of AWE research. This agenda has broadened the scope from the dependability of scoring engines to validating contextualized uses of existing feedback systems by students and teachers in order to discern the intricacies of ecological implementations and to reveal how, when, and why AWE innovations can be effective in practice. The larger portion of research examined the use of AWE with native speakers, focusing mainly on outcomes and error rates and excluding the learning/educational processes. The findings can hardly be considered indisputable. Cumulative evidence suggests that AWE can exert positive impacts, including learners' increased understanding of errors, writing improvement, elevated motivation, and enhanced learner autonomy. At the same time, improvement may not always be significant, and it may be limited to spelling, punctuation, and grammar (Shermis & Burstein, 2013).

In second language (L2) contexts, AWE research is scarcer, often being confined to quantitative and perceptional data. Nevertheless, it suggests that AWE applications hold a promising potential to facilitate improvement in writing accuracy. They appear to be favored more by low-level rather than advanced L2 writers and preferred at earlier stages of drafting and revision, with teacher and peer feedback provided at a later point in the writing process. Weigle (2013) explains that the writing construct as represented in AWE is less of an issue with regards to L2 writers because the learner variables and the foci of instruction are markedly different compared to L1. Form-focused AWE applications seem to be appropriate for contexts where writing needs to be taught as a language skill, with instruction and assessment needing to focus on syntax, morphology, vocabulary use, and paragraph structure, especially at lower levels of proficiency. Emphasis on higher-level concerns is also important, but generally becomes more prominent in intermediate and advanced writing classes.

Specialist discussions and empirical reports, however, are accompanied by mere mentions of pedagogical implications, providing little advice for how to transfer the relatively sporadic suggestions to effective utilization of the feedback applications. When the focus is on language proficiency, "a relatively strong argument can be made for automated scoring and feedback systems if they can be implemented wisely" (Weigle, 2013, p. 39) – "wisely" implies in a grounded and principled way. Generalizing the use of AWE tools across instructional settings, teaching goals, and learning objectives is one of the most dangerous caveats that can vitiate wise utilization. Although AWE research rarely tackles contextual factors explicitly, it tangentially indicates that instructors' implementation choices vary considerably in scope and approach. It is not uncommon for teachers to adopt AWE for scoring or test preparation purposes, reserving very little time for revision and thus largely disregarding these systems' formative feedback capabilities. While alleviating some instructional burdens, such AWE uses can in fact be considered misuse. Other teachers, however, employ AWE to complement more germane types of activities such as pre-writing, writing practice, peer review, revision, and teacher commenting. The outcomes in terms of AWE effectiveness in these cases can be more satisfactory both for students and teachers (Warschauer & Grimes, 2008).

Designing context-based and needs-driven AWE would be a highly plausible option and a significant step forward. Cotos (2014) presents an approach applied to the conceptual design of a corpus and genre-based AWE program called *Intelligent Academic Discourse Evaluator (IADE)*, also providing an

empirical evaluation paradigm for obtaining learner-centered evidence to corroborate design hypotheses from the perspective of second language acquisition and socio-cognitive theories. Developing custom-made applications would be ideal, and, looking into the future, it seems inevitable that AWE will undergo continuous transformations in this vein. In the meantime, the reality is that existing commercial products are increasingly implemented in public schools, community colleges, and universities worldwide. Apparently, AWE applications are "here to stay, and the focus should be on continuing to improve both the human and technological sides of the equation" (Weigle, 2013, p. 50).

Of all the stakeholders involved in the implementation of AWE, teachers and students are arguably the ones who need most guidance. AWE software must be introduced to them with rationales as to why and how particular affordances should be exploited rather than what features they contain. Theoretically, AWE embodies the input, interaction, output, and practice tenets essential for language learning. A unique AWE affordance is transforming learner's own writing into enhanced input, where potential errors are highlighted and specified by the automated feedback. Additionally, exposure to such rather meaningful to the learners input is expected to foster their focus on form and meaning. This is likely to condition learner-computer interaction whereby learners notice negative evidence in their writing and formulate linguistic hypotheses. Inter-personal interaction can be afforded by teacher's feedback, which can often be embedded along with the automated feedback. This affordance brings the human factor into the interaction equation where both computationally detectable errors and more subtle aspects of writing are pointed out. Instant analysis of revisions and iterative feedback provide opportunities for practice leading to modified output and hypothesis verification. These theoretically grounded rationales present AWE as technology capable to mediate humanistic practice to the benefit of developing L2 writers, not as technological disruption that dehumanizes writing instruction.

Certainly, it must be acknowledged that computer algorithms will not reach perfection any time soon, if ever, and provide entirely accurate feedback. Automated analysis algorithms can either fail to identify a problem or categorize something as an error when in fact it is not. The question, however, is not how accurate the feedback can be, but rather how appropriate it may be – given the theoretical expectations and also in view of specific learning and instructional objectives. Imperfect automated feedback may exert both positive impact, as it may trigger learners' noticing of problematic areas (Cotos, 2014), and negative impact, as it may mislead the students into making inappropriate corrections (Weigle, 2013). That is why it is important for teachers to first closely examine the criteria used for the generation of scores and feedback, and then set rational and pragmatic expectations of AWE tools.

It would be advisable, especially for first time users, to get well-acquainted with the capabilities of AWE in order to understand potential strengths and weaknesses and to develop strategies that would compensate for technical limitations. Introducing a smaller set of features at a time could be a plausible approach. For example, teachers may choose to enable only the feedback on grammar and mechanics for a paragraph-writing task and to complement it with their own comments in instances where the tool either misclassifies or misidentifies an error. In this process, teachers would find out which linguistic phenomena are more challenging for automated classification and thus require more elaboration in class. More importantly, they would better understand how to help students process the feedback. For L2 writers, the typically metalinguistic formulation of feedback prompts may not be easy to understand, so teachers may need to explain how to construe the feedback, which would also enhance students' autonomous learning.

Furthermore, while L2 writers generally perceive instant feedback as helpful, it may still fall short because it tends to be fixed and somewhat repetitive in nature. They may feel overwhelmed when seeing multiple types of errors highlighted on the screen and discouraged without guidance for how to prioritize addressing the errors. Grimes and Warschauer (2010) note that their students found it difficult to make sense of the flood of feedback, which is why they supplied clarification handouts. Similarly,

teachers can justifiably advise to focus on most frequent error types (directing students to system-generated holistic reports) and to address errors of word choice and verb use that would likely impede the understanding of ideas (suggesting that students consult relevant resource tools within the AWE system), in addition commenting on concerns that are difficult to automate by embedding their own feedback. Perhaps, teachers could even enhance the mediating role of AWE by developing scenarios of interaction with the feedback and demonstrating how to productively react to it at a given stage in the writing process.

Few AWE discussants address the value of various contextualized help options available to students on demand before, during, and post first-draft stage, which teachers can implement as scaffolding to align theory with classroom practice. For pre-writing, for instance, *Criterion* offers a planning tool containing different templates for planning strategies, and *Folio* offers interactive practice passages for revision and engaging animated tutorials. Scaffolding features accessible to the students during writing and revision abound in AWE programs. Both *Criterion* and *MyAccess!* generate multilingual feedback for L2 writers, and *WriteToLearn* offers word translation and uses text-to-speech technologies that articulate written text – such features can be particularly helpful as comprehension-facilitating scaffolds for lower-level learners. *Criterion*'s context-sensitive handbook offers definitions and examples of correct and incorrect use. The 'just-in-time' writing assistance of *MyAccess!* includes a bank of appropriate vocabulary use, an editor with suggestions for correcting highlighted errors, and a writing coach with remediation activities for specific writing traits.

Undoubtedly, effective AWE implementation is contingent on teacher support and principled choices tied to specific objectives. The objectives, in many cases, are dictated by institutional goals and administrative strategies, in view of which teachers may be strongly encouraged to use the score reporting features that can yield performance reports at student, class, school, and in some cases even district level. The use of such features requires prudence and forethought. Similar to psychometric research on AES, automated scores have been juxtaposed with teacher grades, yielding much less encouraging correlations compared to the relatively high indices between scoring engines and human raters. Therefore, teachers need to be cautious about using automated scores for grading and opt to exploit them for diagnostic or pre-assessment benchmark purposes instead. While, reports generated at the level of student history and group comparison can provide teachers with the necessary information for curriculum planning, using them as measures of achievement at institutional level would be quite a feeble extrapolation that would fuel the AWE debate.

SEE ALSO:

Automated Writing Assessment; CALL and Feedback; Input Enhancement; Noticing Hypothesis; Scaffolding Technique; Natural Language Processing and ICALL

References

Cheville, J. (2004). Automated scoring technologies and the rising influence of error. *English Journal, 93*(4), 47–52.

Cotos, E. (2014). *Genre-based Automated Writing Evaluation for L2 Research Writing: From Design to Evaluation and Enhancement*. New York, NY: Palgrave Macmillan.

Grimes, D., & Warschauer, M. (2010). Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation. *Journal of Technology, Learning, and Assessment, 8*(6), 1-43.

Perelman, L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). In C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (Eds.), *International advances in writing research: Cultures, places, measures* (pp. 121-131). Fort Collins, Colorado: WAC Clearinghouse/Anderson.

Shermis, M.D. & Burstein, J.C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Shermis, M.D. & Burstein, J.C. (2013). (Eds). *Handbook of Automated essay Evaluation: Current applications and new directions*. Routledge, New York.

Warschauer, M., & Grimes, G. (2008). Automated Writing Assessment in the Classroom. *Pedagogies: An International Journal, 3*, 22-36.

Weigle, S.C. (2013). English as a second language writing and Automated Essay Evaluation. In M.D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current applications and new directions* (pp. 36-54). Routledge, New York.

Further Readings

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*, 7–24.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research, 10*(2), 1-24.