Automated Scoring of Writing

Stephanie Link and Svetlana Koltovskaia

Abstract For decades, automated essay scoring (AES) has operated behind the scenes of major standardized writing assessments to provide summative scores of students' writing proficiency (Dikli in J Technol Learn Assess 5(1), 2006). Today, AES systems are increasingly used in low-stakes assessment contexts and as a component of instructional tools in writing classrooms. Despite substantial debate regarding their use, including concerns about writing construct representation (Condon in Assess Writ 18:100–108, 2013; Deane in Assess Writ 18:7–24, 2013), AES has attracted the attention of school administrators, educators, testing companies, and researchers and is now commonly used in an attempt to reduce human efforts and improve consistency issues in assessing writing (Ramesh and Sanampudi in Artif Intell Rev 55:2495–2527, 2021). This chapter introduces the affordances and constraints of AES for writing assessment, surveys research on AES effectiveness in classroom practice, and emphasizes implications for writing theory and practice.

Keywords Automated essay scoring · Summative assessment

1 Overview

Automated essay scoring (AES) is used internationally to rapidly assess writing and provide summative holistic scores and score descriptors for formal and informal assessments. The ease of using AES for response to writing is especially attractive for large-scale essay evaluation, providing also a low-cost supplement to human scoring and feedback provision. Additionally, intended benefits of AES include the elimination of human bias, such as rater fatigue, expertise, severity/leniency, inconsistency,

S. Link (✉)
Oklahoma State University, 205 Morrill Hall, Stillwater, OK 74078, USA
e-mail: steph.link@okstate.edu

S. Koltovskaia
Department of Languages and Literature, Northeastern State University, Tahlequah, OK 74464, USA
e-mail: koltovsk@nsuok.edu

O. Kruse et al. (eds.), *Digital Writing Technologies in Higher Education*,
https://doi.org/10.1007/978-3-031-36033-6_21

and Halo effect. While AES developers also commonly suggest that their engines perform as reliably as human scorers (e.g., Burstein & Chodorow, 2010; Riordan et al., 2017; Rudner et al., 2006), AES is not free of critique. Automated scoring is frequently under scrutiny for use with university-level composition students in the United States (Condon, 2013) and second language writers (Crusan, 2010), with some writing practitioners discouraging its replacement of adequate literacy education because of its inability to evaluate meaning from a humanistic, socially-situated perspective (Deane, 2013; NCTE, 2013). AES also suffers from biases, such as imperfections in the quality and representation of training data to develop the systems and inform feedback generation. These biases question the fairness of AES (Loukina et al., 2019), especially if scores are modeled based on data that does not adequately represent a user population—a particular concern for use of AES with minoritized populations.

Despite reservations, the utility of AES in writing practices has increased significantly in recent years (Ramesh & Sanampudi, 2021), partially due to its integration into classroom-based tools (see Cotos, "Automated Feedback on Writing" for a review of automated writing evaluation). Thus, the affordances of AES for language testing are now readily available to writing practitioners and researchers, and the time is ripe for better understanding its potential impact on the pedagogical approaches to writing studies by first better understanding the history that drives AES development.

Dating back to the 1960s, AES started with the advent of Project Essay Grade (Page, 1966). Since then, automated scoring has advanced into leading technologies, including e-rater by the Educational Testing Service (ETS) (Attali & Burstein, 2006), Intelligent Essay Assessor (IEA) by Knowledge Analysis Technologies (Landauer et al., 2003), Intellimetric by Vantage Learning (Elliot, 2003), and a large number of prospective newcomers (e.g., Nguyen & Dery, 2016; Riordan et al., 2017). These AES engines are used for tests like the Test of English as a Foreign Language (TOEFL iBT), Graduate Management Admissions Test (GMAT), and the Pearson Test of English (PTE). In such tests, AES researchers not only found the scores reliable, but some argued that they also allowed for reproducibility, tractability, consistency, objectivity, item specification, granularity, and efficiency (William et al., 1999), characteristics that human raters can lack (Williamson et al., 2012).

The immediate AES response to writing is without much question a salient feature of automated scoring for testing contexts. However, research on classroom-based implementation has suggested that instructors can utilize the AES feedback to flag students' writing that requires teachers' special attention (Li et al., 2014), highlighting its potential for constructing individual development plans or conducting analysis of students' writing needs. AES also provides constant, individualized feedback to lighten instructors' feedback load (Kellogg et al., 2010), enhance student autonomy (Wang et al., 2013), and stimulate editing and revision (Li et al., 2014).

2 Core Idea of the Technology

Automated essay scoring involves automatic assessment of a students' written work, usually in response to a writing prompt. This assessment generally includes (1) a holistic score of students' performance, knowledge, and/or skill and (2) a score descriptor on how the student can improve the text. For example, e-rater by ETS (2013) scores essays on a scale from 0 to 6. A score of 6 may include the following feedback:

> Score of 6: Excellent
> Your essay
> Looks at the topic from a number of angles and responds to all aspects.
> Responds thoughtfully and insightfully to the issues in the topic.
> Develops with a superior structure and apt reasons or examples.
> Uses sentence styles and language that have impact and energy.
> Demonstrates that you know the mechanics of correct sentence structure.

AES engine developers over the years have undertaken a core goal of making the assessment of writing accurate, unbiased, and fair (Madnani & Cahill, 2018). The differences in score generation, however, are stark given the variation in philosophical foundations, intended purposes, extraction of features for scoring writing, and criteria used to test the systems (Yang et al., 2002). To this end, it is important to understand the prescribed use of automated systems so that they are not implemented inappropriately. For instance, if a system is meant to measure students' writing proficiency, the system should not be used to assess students' aptitude. Thus, scoring models for developing AES engines are valuable and effective in distinct ways and for their specific purposes.

Because each engine may be designed to assess different levels, genres, and/or skills of writing, developers utilize different natural language processing (NLP) techniques for establishing construct validity, or the extent to which an AES scoring engine measures what it intends to measure—a common concern for AES critics (Condon, 2013; Perelman, 2014, 2020). NLP helps computers understand human input (text and speech) by starting with human and/or computer analysis of textual features so that a computer can process the textual input and offer reliable output (e.g., a holistic score and score descriptor) on new text. These features may include statistical features (e.g., essay length, word co-occurrences also known as n-grams), style-based features (e.g., sentence structure, grammar, part-of-speech), and content-based features (e.g., cohesion, semantics, prompt relevance) (see Ramesh & Sanampudi, 2021, for an overview of features). Construct validity should thus be interpreted in relation to feature extraction of a given AES system to adequately appreciate (or challenge) the capabilities that system offers writing studies.

In addition to a focus on a variety of textual features, AES developers have utilized varied machine learning (ML) techniques to establish construct validity and efficient score modeling. Machine learning is a category of artificial intelligence (AI) that helps computers recognize patterns in data and continuously learn from the data to

make accurate holistic score predictions and adjustments without further programming (IBM, 2020). Early AES research utilized standard multiple regression analysis to predict holistic scores based on a set of rater-defined textual features. This approach was utilized in the early 1960s for developing Project Essay Grade by Page (1966), but it has been criticized for its bias in favor of longer texts (Hearst, 2000) and its ignorance towards content and domain knowledge (Ramesh & Sanampudi, 2021).

In subsequent years, classification models, such as the bag of words approach (BOW), were common (e.g., Chen et al., 2010; Leacock & Chodorow, 2003). BOW models extract features in writing using NLP by counting the occurrences and co-occurrences of words within and across texts. Texts with multiple shared word strings are classified into similar holistic score categories (e.g., low, medium, high) (Chen et al., 2010; Zhang et al., 2010). E-rater by ETS is a good example of this approach. The aforementioned approaches are human-labor intensive. Latent semantic analysis (LSA) is advantageous in this regard; it is also strong in evaluating semantics. In LSA, the semantic representation of a text is compared to the semantic representation of other similarly scored responses. This analysis is done by training the computer on specific corpora that mimics a given writing prompt. Landauer et al. (2003) used LSA in Intelligent Essay Grade.

Advances in NLP and progress in ML have motivated AES researchers to move away from statistical regression-based modeling and classification approaches to advanced models involving neural network approaches (Dong et al., 2017; Kumar & Boulanger, 2020; Riordan et al., 2017). To develop these AES models, data undergoes a process of supervised learning, where the computer is provided with labeled data that enables it to produce a score as a human would. The supervised learning process often starts with a training set—a large corpus of representative, unbiased writing that is typically human- or auto-coded for specific linguistic features with each text receiving a holistic score. Models are then generated to teach a computer to identify and extract these features and provide a holistic score that correlates with the human rating. The models are evaluated on a testing set that the computer has never seen previously. Accuracy of algorithms is then evaluated by using testing set scores and human scores to determine human–computer consistency and reliability. Common evaluations are quadrated weighted kappa, Mean Absolute Error, and Pearson Correlation Coefficient.

Once accuracy results meet an industry standard (Powers et al., 2015), which varies across disciplines (Weigle, 2013), the algorithms are made public through user-friendly interfaces for testing contexts (i.e., to provide summative feedback, formal assessments to assess students' performance or proficiency) and direct classroom use (i.e., informal assessments to improve students' learning). For the classroom, teachers should be active in evaluating the feedback to determine whether it is reasonably accurate in assessing a learning goal, does not lead students away from the goal, and encourages students to engage in different ways with their text and/or the course content. Effective evaluation of AES should start with an awareness of AES affordances that can impact writing practice and then continue with the training of students in the utility of these affordances.

3 Functional Specifications

The overall functionality of AES for classroom use is to provide summative assessment of writing quality. AES accomplishes this through two key affordances: a holistic score and score descriptor.

Holistic score: The summative score provides an overall, generic assessment of writing quality. For example, Grammarly provides a holistic score or "performance" score out of 100%. The score represents the quality of writing (as determined by features, such as word count, readability statistics, vocabulary usage). If a student receives a score below 60–70%, this means that it could be understood by a reader who has a 9th grade education. For the text to be readable by 80% of English speakers, Grammarly suggests getting at least 60–70%.

Score descriptor: The holistic score is typically accompanied by a descriptor that indicates what the score represents. This characterization of the score meaning can be used to interpret the feedback, evaluate the feedback, and make decisions regarding editing and revising.

That is, these key affordances can be utilized to complete several main activities.

Interpreting feedback: Once students receive the holistic score along with the descriptor, they should interpret the score. Information provided for adequate score interpretation varies across AES systems, so students may need help in interpreting the meaning of this feedback.

Evaluating feedback: After interpreting the score and the descriptor, students need to think critically about how the feedback applies to their writing. That is, students need to determine whether the computer feedback is an adequate representation of their writing weaknesses. Evaluating feedback thus entails noticing the gap or problem found in one's own writing and becoming consciously aware of how the feedback might be used to increase the quality of writing through self-editing (Ferris, 2011).

Making a decision about action: Once students evaluate their writing based on a given score and descriptor, they then need to decide whether to address the issues highlighted in the descriptor or seek additional feedback. Making and executing a revision plan can ensure that the student is being critical towards the feedback rather than accepting it outright.

Revising/editing: The student then revises the paper and resubmits it to the system to see if the score improves—an indicator of higher quality writing. If needed, the student can repeat the above actions or move on to editing of surface-level writing concerns.

4 Research on AES

AES research can be categorized along two lines: system-centric research that evaluates the system itself and user-centric research that evaluates use/impact of a system on learning. From a system-centric perspective, various studies have been conducted to validate AES-system-generated scores for the testing context. The majority have focused on reliability, or the extent to which results can be considered consistent or stable (Brown, 2005). They often evaluate reliability based on agreement between human and computer scoring (e.g., Burstein & Chodorow, 1999; Elliot, 2003; Streeter et al., 2011). (See Table 1 for a summary of reliability statistics from three major AES developers.)

The process of establishing validity should not start and stop with inter-coder reliability; however, automated scoring presents some distinctive validity challenges, such as "the potential to under- or misrepresent the construct of interest, vulnerability to cheating, impact on examinee behavior, and score users' interpretation and use of scores" (Williamson et al., 2012, p. 3). Thus, some researchers have also demonstrated reliability by using alternative measures, such as the association with independent measures (Attali et al., 2010) and the generalizability of scores (Attali et al., 2010). Others have gone a step further and suggested a unified approach to AES validation (Weigle, 2013, Williamson et al., 2012). In general, results reveal promising developments in AES with modest correlations between AES and external criteria, such as independent proficiency assessments (Attali et al., 2010; Powers et al., 2015, suggesting that automated scores can relate in a similar manner to select assessment criteria and that both have the potential to reflect similar constructs, although results across AES systems can vary, and not all data are readily available to the public.

While much research has focused on reliability of AES, little is known about the quality of holistic scores in testing or classroom contexts as well as teachers' and students' use and perceptions of automatically generated scores. In a testing

Table 1 Summary of human–computer reliability studies from three top developers

AES system	Testing context[a]	Prompt types	Human–Computer Reliability	Study
e-rater	GRE TOEFL iBT	Argument and issues prompts	Weighted Kappa 0.70–0.78 Pearson's r 0.70–0.80	Attali et al. (2010)
IntelliMetric	GMAT	Argument and issues prompts	Pearson's r 0.80–0.84	Rudner et al. (2006)
Intelligent Essay Assessor	PTE	Argument, issues, and narrative prompts	Pearson's r 0.88–0.91	Streeter et al. (2011)

Note [a]GRE = Graduate Record Examination
TOEFL = Test of English as a Foreign Language internet-based test
GMAT = Graduate Management Admission Test
PTE = Pearson Test of English

context, James (2006) compared the IntelliMetric scores of the ACCUPLACER OnLine WritePlacer Plus test to the scores of "untrained" faculty raters. Results revealed a relatively high level of correspondence between the two. In a similar study with a group of developmental writing students in a two-year college in South Texas, Wang and Brown (2007) found that ACCUPLACER's overall holistic mean score showed significant difference between IntelliMetric and human raters, indicating that IntelliMetric tends to assign higher scores than human raters do. Li et al. (2014) investigated the correlation between Criterion's numeric scores with the English as a second language instructors' numeric grades and analytic ratings for classroom-based assessment. The results showed low to moderate positive correlations between Criterion's scores and instructors' scores and analytic ratings. Taken together, these studies suggest limited continuity of findings on AES reliability across tools.

Results of multiple studies demonstrate varied uses for holistic scores and varied teachers' and students' perceptions toward the scores. For example, Li et al. (2014) found that Criterion's holistic scores in the English as a second language classroom were used in three ways. First, instructors used the scores as a forewarning. That is, the scores alerted instructors to problematic writing. Second, the scores were used as a pre-submission benchmark. That is, the students were required to obtain a certain score before submitting a final draft to their teacher. Finally, Criterion's scores were utilized as an assessment tool—scores were part of course grading. Similar findings were reported in Chen and Cheng's (2008) study that focused on EFL Tawainese teachers' and students' use and perception of My Access! While one teacher used My Access! as a pre-submission benchmark, the other used it for both formative and summative assessment, heavily relying on the scores to assessing writing performance. The third teacher did not make My Access! a requirement and asked the students to use it if they needed to.

In terms of teachers' perceptions of holistic scores, holistic scores seem to be motivators for promoting student revision (Li et al. 2014; Scharber et al., 2008) although a few teachers in Maeng (2010) commented that the score caused some stress albeit was still helpful for facilitating the feedback process (i.e., for providing sample writing and revising). Teachers also tend to have mixed confidence in holistic scores (Chen & Cheng, 2008; Li et al, 2014). For example, in Li et al.'s (2014) study, English as a second language instructors had high trust in Criterion's low holistic scores as the essays Criterion scored low were, in fact, poor essays. However, instructors possessed low levels of trust when Criterion assigned high scores to writing as instructors judged such writing lower.

Students also tend to have low trust in holistic scores (Chen & Cheng, 2008; Scharber et al., 2008). For example, Chen and Cheng (2008) found that EFL Taiwanese students' low level of trust in holistic scores was influenced by teachers' low level of trust in the scores as well as discrepancies in teachers' scores and holistic scores of My Access! that students noticed. Similar findings were reported in Scharber et al.'s (2008) study that focused on Educational Theory into Practice Software's (ETIPS) automated scorer implemented in a post-baccalaureate program at a large public Midwestern US university. The students in their study experienced negative emotions due to discrepancies in teachers' and ETIPS' holistic scores. ETIPS

scores were one point lower than teachers' scores. Additionally, the students found holistic scores with the short descriptor insufficient in guiding them as to how to actually improve their essays.

5 Implications of This Technology for Writing Theory and Practice

The rapid advancement of NLP and ML approaches to automated scoring lends well to theoretical contributions that help to (re-)define traditional notions of how learning takes place and the phenomena that underscores language development. Social- and cognitive-based theories to writing studies can be expanded with the integration of AES technology by offering new, socially-situated learning opportunities in online environments that can impact how students respond to feedback. These digitally-rich learning opportunities can thus significantly impact the writing process, offering a new mode of feedback that can be meaningful, constant, timely, and manageable while addressing individual learner needs. From a traditional pen-and-paper approach, these benefits are known to contribute significantly to writing accuracy (Hartshorn et al., 2010), and so the addition of rapid technology has the potential to add new knowledge to writing development research.

AES research can also contribute to practice. Due to its instantaneous nature, AES holistic scores could be used for placement purposes (e.g., by using ACCUPLACER) at schools, colleges, and universities. However, relying on the AES holistic score alone may not be adequate. Therefore, just like in large-scale tests, it is important that students' writing is double-rated to enhance reliability, with a third rater used if there is a discrepancy in AES holistic score and a human rater's score. Similarly, AES holistic scores could be used for diagnostic assessment. Diagnostic assessment is given prior to or at the start of the semester/course to get information about students' language proficiency as well as their strengths and weaknesses in writing. Finally, AES scoring could be used for summative classroom assessment. For example, teachers could use AES scores as a pre-submission benchmark and require students to revise their essays until they get a predetermined score, or teachers could use the AES score for partial (rather than sole) assessment of goal attainment (Li et al., 2014; Weigle, 2013). Overall, in order to avoid pitfalls such as students focusing too intensively on obtaining high scores without actually improving their writing skills, teachers and students need to be trained or seek training on the different merits and demerits of a selected AES scoring system.

6 Concluding Remarks

While traditional approaches to written corrective feedback are still leading writing studies research, the ever-changing digitalization of the writing process shines light on new opportunities for enhancing the nature of feedback provision. The evolution of AI will undoubtedly expand the affordances of AES so that writing in digital spaces can be supplemented by computer-based feedback that is increasingly accurate and reliable. For now, these technologies are only foregrounding what can come from technological advancements, and in the meantime, it is the task of researchers and practitioners to cast a critical eye while also remaining open to the potential for AES technologies to promote autonomous, lifelong learning and writing development.

7 Tool List

List of well-known Automated Essay Scoring (AES) Tools

N	Tool	Description	Suggested use	Reference
1	**E-rater** in Criterion (https://criterion.ets.org/criterion/default.aspx) and Turnitin (https://www.turnitin.com/)	E-rater was developed by Educational Testing Service (ETS) to identify features related to writing proficiency in student essays	The suggested use is with middle school to high school students with writing prompts available for first- and second-year university students	Attali and Burstein (2006)
2	**Intellimetric** (https://www.intellimetric.com/direct)	Intellimetric that was developed by Vantage Learning is a web-based tool capable of scoring short and long writing pieces in more than 20 languages (e.g., English varieties, Bahasa Malaysia, Chinese, Turkish, and Spanish)	Although marketed for all aged-writers, most research using Intellimetric is found to successfully assess writing of middle-schoolers (about ages 11–13) and those seeking writing placement using the accompanying technology the ACCUPLACER OnLine WritePlacer Plus test, a standardized placement test that measures writing proficiency of entry-level college students (https://accuplacer.collegeboard.org/)	Elliot (2003)

(continued)

(continued)

N	Tool	Description	Suggested use	Reference
3	**Intelligent Essay Assessor (IEA)** (https://www.pearsonassessments.com/)	IEA uses knowledge analysis technologies (KAT) engine and is available in Pearson's WritetoLearn web-based tool	IEA is intended for grades 4–12. This technology can assess English, Spanish, and Chinese writers	Landauer et al. (2003)
4	**Educational Theory into Practice Software (ETIPS)** (http://www.etips.info/)	ETIPS is an online learning environment that was developed in 2003. Its AES engine is built using a Bayseian model for essay scoring. It is noteworthy that ETIPS AES does not score essays that "deal with other than ETIPS case-specific questions and topics" (Scharber et al., 2008, p. 9)	Its intended audience are pre-service teachers preparing for technology implementation in their classrooms. Its embedded assessment feature is designed for K-12 students	Dexter (2007)

References

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *Journal of Technology, Learning, and Assessment, 10*(3). http://www.jtla.org

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment, 4*(3), 1–30.

Brown, J. D. (2005). *Testing in language programs. A comprehensive guide to English language assessment*. McGraw Hill.

Burstein, J., & Chodorow, M. (1999). *Automated essay scoring for nonnative English speakers.* Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing. http://www.ets.org/Media/Research/pdf/erater_acl99rev.pdf

Burstein, J., & Chodorow, M. (2010). Progress and new directions in technology for automated essay evaluation. In R. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (2nd ed., pp. 487–497). Oxford University Press.

Chen, C., & Cheng, W. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology, 12*(2), 94–112.

Chen, Y. Y., Liu, C. L., Chang, T. H., & Lee, C. H. (2010). An unsupervised automated essay scoring system. *IEEE Intelligent Systems, 25*(5), 61–67. https://doi.org/10.1109/MIS.2010.3

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing, 18*, 100–108. https://doi.org/10.1016/j.asw.2012.11.001

Crusan, D. (2010). *Assessment in the second language writing classroom*. University of Michigan Press.
Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*, 7–24.
Dexter, S. (2007). Educational theory into practice software. In D. Gibson, C. Aldrich, & M. Prensky (Eds.), *Games and simulations in online learning: Research and development frameworks* (pp. 223–238). IGI Global. https://doi.org/10.4018/978-1-59904-304-3.ch011
Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment, 5*(1). https://ejournals.bc.edu/index.php/jtla/article/view/1640
Dong, F., Zhang, Y., & Yang, J. (2017). *Attention-based recurrent convolutional neural network for automatic essay scoring*. Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). https://aclanthology.org/K17-1017.pdf
Elliot, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis & J. C. Burstein (Eds.), *Automatic essay scoring: A cross-disciplinary perspective* (pp. 71–86). Lawrence Erlbaum Associates.
ETS. (2013). *Criterion scoring guide*. Retrieved September 27, 2013, from http://www.ets.org/Media/Products/Criterion/topics/co-1s.htm
Ferris, D. R. (2011). *Treatment of errors in second language student writing* (2nd ed.). The University of Michigan Press.
Hartshorn, K. J., Evans, N. W., Merrill, P. F., Sudweeks, R. R., Strong-Krause, D., & Anderson, N. J. (2010). Effects of dynamic corrective feedback on ESL writing accuracy. *TESOL Quarterly, 44*, 84–109.
Hearst, M. (2000). The debate on automated essay grading. *IEEE Intelligent Systems and their Applications, 15*(5), 22–37. https://doi.org/10.1109/5254.889104
IBM. (2020). *Machine learning*. IBM Cloud Education. https://www.ibm.com/cloud/learn/machine-learning
James, C. (2006). Validating a computerized scoring system for assessing writing and placing students in composition courses. *Assessing Writing, 11*(3), 167–178.
Kellogg, R., Whiteford, A., & Quinlan, T. (2010). Does automated feedback help students learn to write? *Journal of Educational Computing Research, 42*, 173–196.
Kumar, V., & Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education (Lausanne), 5*. https://doi.org/10.3389/feduc.2020.572367
Landauer, T. K., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education, 10*(3), 295–308.
Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37*, 389–405.
Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System, 44*, 66–78. https://doi.org/10.1016/j.system.2014.02.007
Loukina, A., et al. (2019). *The many dimensions of algorithmic fairness in educational applications*. BEA@ACL.
Madnani, N., & Cahill, A. (2018). *Automated scoring: Beyond natural language processing*. COLING.
Maeng, U. (2010). The effect and teachers' perception of using an automated essay scoring system in L2 writing. *English Language and Linguistics, 16*(1), 247–275.
NCTE. (2013, April 20). *NCTE position statement on machine scoring*. National Council of Teachers of English. https://ncte.org/statement/machine_scoring/
Nguyen, H., & Dery, L. (2016). *Neural networks for automated essay grading* (pp. 1–11). CS224d Stanford Reports.
Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48*, 238–243.
Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing, 21*, 104–111.

Perelman, L. (2020). The BABEL generator and E-rater: 21st century writing constructs and automated essay scoring (AES). *Journal of Writing Assessment, 13*(1).

Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the "gold standard." *Applied Measurement in Education, 28*(2), 130–142. https://doi.org/10.1080/08957347.2014.1002920

Ramesh, D., & Sanampudi, S. K. (2021). An automated essay scoring systems: A systematic literature review. *The Artificial Intelligence Review, 55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Riordan, B., Horbach, A., Cahill, A., Zesch, T., & Lee, C. M. (2017). *Investigating neural architectures for short answer scoring*. Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. https://aclanthology.org/W17-5017.pdf

Rudner, L., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetricTM essay scoring system. *Journal of Technology, Learning, and Assessment, 4*(4). http://escholarship.bc.edu/ojs/index.php/jtla/article/view/1651/1493

Scharber, C., Dexter, S., & Riedel, E. (2008). Students' experiences with an automated essay scorer. *Journal of Technology, Learning and Assessment, 7*(1), 1–45. https://ejournals.bc.edu/index.php/jtla/article/view/1628

Streeter, L., Bernstein, J., Foltz, P., & DeLand, D. (2011). *Pearson's automated scoring of writing, speaking, and mathematics.* White Paper. http://images.pearsonassessments.com/images/tmrs/PearsonsAutomatedScoringofWritingSpeakingandMathematics.pdf

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment, 6*(2). http://www.jtla.org

Wang, Y., Shang, H., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning, 26*(3), 1–24.

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

William, D. M., Bejar, I. I., & Hone, A. S. (1999). 'Mental model' comparison of automated and human scoring. *Journal of Educational Measurement, 35*(2), 158–184.

Williamson, D., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15*(4), 391–412. https://doi.org/10.1207/S15324818AME1504_04

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics, 1*, 43–52.

Stephanie Link, PhD, is Associate Professor of Applied Linguistics and Director of Graduate Studies at Oklahoma State University in Stillwater, Oklahoma, USA. Her work in computer assisted language learning focuses on leveraging textual mining techniques, natural language processing, and genre theory to support second language writers and scientific writers in writing for publication.

Svetlana Koltovskaia, PhD, is Assistant Professor of English and director of the ESL Academy at Northeastern State University, Tahlequah, Oklahoma, USA. Her research centers around L2 writing, computer-assisted language learning, and L2 assessment.