Automated essay scoring (AES) systems: Opportunities and challenges for open and distance education

John Y. H. Bai[1], Olaf Zawacki-Richter[1], Aras Bozkurt[2], Kyungmee Lee[3], Mik Fanguy[3], Berrin Cefa Sari[1], and Victoria I. Marín[4]

[1]Carl von Ossietzky University of Oldenburg, Germany

[2]Anadolu University, Turkey

[3]Lancaster University, United Kingdom

[4]Universitat de Lleida, Spain

Artificial intelligence applications in Education (AIEd) are a growing research topic. Zawacki-Richter et al. (2019) systematically reviewed published AIEd articles and identified four general areas of application: 1) profiling and prediction, 2) assessment and evaluation, 3) adaptive systems and personalisation, and 4) intelligent tutoring systems. The range of different applications demonstrates the potential of AIEd to transform educational practices. To guide successful development and implementation, we need to understand how AIEd works and ensure multi-stakeholder dialogue throughout its lifecycle (Hu et al., 2019; IEEE, 2019; UNESCO, 2021). Thus, the present study focuses on one specific type of AIEd falling under "assessment and evaluation" in Zawacki-Richter et al.'s review: automated essay scoring (AES). AES systems could significantly reduce teacher workload, especially in large Open and Distance Learning courses. This paper extends and synthesizes a corpus of articles collected from Zawacki-Richter et al.'s original review and a recently updated replication to understand how AES systems work.

Researchers (e.g., Gierl et al., 2014; McNamara et al., 2015; Wilson et al., 2021) trace AES systems back to Page and colleagues' Project Essay Grade (PEG; see Page, 1966). While technological developments have increased the complexity and accuracy of AES, core concepts in Page's work remain relevant. Specifically, Page distinguishes between intrinsic versus approximate variables, termed *trin* and *prox*. Trins are variables of interest human raters may use to score an essay (e.g., sentence structure, organization, word choice, etc.). However, trins cannot be directly measured. Thus, AES systems instead use proxes – variables that correlate with the trins they are chosen to approximate (e.g., the relative frequencies of longer vs. shorter, common vs. uncommon words used to approximate "word choice").

Proxes are the *features* of the texts that AES systems evaluate. Multiple articles distinguish between "surface-level features" (e.g., spelling mistakes, sentence length, essay length, etc.) and "deeper features" that reflect semantic or rhetoric dimensions of an essay (Latifi & Gierl, 2021; McNamara et al., 2015; Raković et al., 2021; Yang et al., 2019). Common feature-extraction tools include Coh-Metrix (Graesser et al., 2004), which includes indices of over 100 features, and Latent Semantic Analysis (Foltz, 1996), which uses relative word distributions as proxies for related semantic content.

AES systems map the extracted features onto human-rated essay scores. This often involves supervised machine learning, wherein a model is trained with a set of human-scored essays, and then predicts the scores of another set of essays by comparing their similarity with essays in the training set (see Gierl et al., 2014; Ma & Slater, 2015; Mayfield & Rose, 2013). Statistical techniques (e.g., linear and logistic regression) can also be used to predict scores; however, machine learning models such as support vector machines (SVM) and random forests are growing more common in recent research (see e.g., Ifenthaler, in press, for a historical perspective).

Comparing the patterns of proxes in higher versus lower human-scored essays to grade essays has received a fair amount of criticism (e.g., Ericcson & Haswell 2006; Perelman, 2012). Perelman and colleagues' Basic Automatic BS Essay Language (BABEL) Generator demonstrates the current limits of AES by producing "gibberish" texts that are assigned high scores by multiple AES systems (Perelman, 2020). These texts are adversarial examples (Kumar et al., 2020) that exploit the gap between proxes and trins. Therefore, we investigated whether recent AES articles critically discussed the limitations of AES systems while exploring general trends in the updated review.

Method

This review is part of a larger project that aims to update and replicate the systematic review by Zawacki-Richter et al. (2019) who synthesized 146 studies published between 2007 and October 2018. The original review identified 36 articles in assessment and evaluation, from which eight were identified as strictly about AES.

To update the corpus of articles, we used a replicable search strategy and explicit inclusion and exclusion criteria (Gough et al., 2017; Zawacki-Richter et al., 2020). In November 2021, the search was conducted using an identical search string as Zawacki-Richter et al. (2019; Table 1) in the Web of Science, Scopus, and Education Source databases. A total of 2,592 references were identified, of which 1,822 remained after screening on title and abstracts and removing duplicates. After screening based on the inclusion and exclusion criteria (Table 2), the full texts of 645 studies were retrieved and reviewed; 428 studies were included in the corpus for the synthesis and coded in four areas of applications (Zawacki-Richter et al., 2019), with 99 articles in assessment and evaluation and 13 as strictly AES.

Table 1

Initial Search String

Topic	Search terms
Artificial intelligence	"artificial intelligence" OR "machine intelligence" OR "intelligent support" OR "intelligent virtual reality" OR "chat bot*" OR "machine learning" OR "automated tutor" OR "personal tutor*" OR "intelligent agent*" OR "expert system" OR "neural network" OR "natural language processing"
AND	
Education level	"higher education" OR college* OR undergrad* OR graduate OR postgrad* OR "K-12" OR kindergarten* OR "corporate training*" OR "professional training*" OR "primary school*" OR "middle school*" OR "high school*" OR "elementary school*" OR "vocational education" OR "adult education"
AND	
Learning setting	learn* OR student*

Table 2

Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
Published November 2018 – October 2021	Published November 2018
English language	Not in English
Empirical, primary research	Not primary research (e.g., review)
Indexed in Web of Science, Scopus or EBSCO Education Source	Not a journal article No artificial intelligence
Artificial intelligence use in education	No learning setting

A limitation to the present search strategy is that the search string was designed to capture a broad overview of AIEd rather than specific articles on AES. Therefore, the search string did not contain all combinations of different terms associated with AES (e.g., computer-assisted scoring [CAS], automated writing evaluation [AWE], automated essay evaluation [AEE]). In contrast to AES systems that simply assign scores, McNamara et al. (2015) note that AWE commonly refers to systems that also provide formative feedback. While the present study uses AES as the generic term, future systematic reviews could incorporate diverse terms in the search string.

Results

The present paper synthesizes 21 AES articles: eight from the original and 13 from the updated review. Given the critiques of AES systems, we coded the articles for critical discussions of AES systems beyond the specific limitations of their research design.

Twelve articles included critical discussions on AES systems (Alsanie et al., 2021; Aluthman, 2016; Dikli, 2010; García-Gorrostieta et al., 2018; Hanlon et al., 2021; Knight et al., 2018; Lu, 2019; McNamara et al., 2015; Perin & Lauterbach, 2018; Rupp et al., 2019; Wang, 2020; Wilson et al., 2021). Several articles discussed the inability of AES systems to understand complex texts (Lu, 2019; Perin & Lauterbach, 2018). Another major theme addressed the quality of feedback. For example, Wilson et al. (2021) noted that, despite its accuracy, AWE feedback did not consider the student's effort or characteristics (e.g., students may receive feedback on skills that were not covered in the curriculum). Similarly, Dikli (2010) noted that feedback from teachers, unlike AWE, could track student development by referring "to errors students made in the previous drafts" (p. 119). García-Gorrostieta et al. (2018, p. 1225) commented on the need "to work on the textual feedback, turning it more dynamic and personalized". Other studies also noted deficiencies in the feedback for non-native or beginner level students (Aluthman, 2016; Dikli, 2010).

We divided the corpus into 11 articles that evaluated the performance of AES systems (e.g., Hanlon et al., 2021; Ma & Slater, 2015; McNamara et al., 2015; Raković et al., 2021; Rupp et al., 2019; Yang et al., 2019), nine that evaluated user experiences (e.g., Dikli, 2010; García-Gorrostieta et al., 2018; Lu, 2019; Nazari et al., 2021; Wilson et al., 2021; Wang, 2020), and one that evaluated both (Knight et al., 2018).

Evaluations of AES Systems

The general structure of articles within this section consisted of 1) collecting a set of essays, 2) scoring by expert/trained raters, 3) training an AES system with some of the pre-scored essays, and 4) applying the trained system to score the remaining essays. The performance of the AES is then evaluated by comparing AES-generated scores (the *predicted* scores) against the "golden standard" of human-assigned scores (the *actual* scores).

Different articles used different agreement measures. The simplest measures are the correlation between human and machine scoring (Pearson's *r*) and the exact accuracy (i.e., the percentage of cases when both human and machine agree on the exact score). Other metrics were derived from a confusion matrix to reflect the patterns of correct and incorrect predictions (see Mayfield & Rose, 2013, for an introduction to confusion matrices). From the counts of predicted scores in the confusion matrix cells, researchers can compute metrics like the percentage of adjacent accuracy (e.g., when the actual score is 3, and the predicted score is either 2, 3, or 4), the F1 score (the harmonic mean of precision and recall scores), and the quadratic weighted kappa (*QWK*) which differentially penalizes greater deviations from the actual score (e.g., with an actual score of 3, a predicted score of 6 is penalized more than a predicted score of 5). While the specifics of the measures differ, F1 and kappa values closer to 1.0 indicate closer agreement between actual and predicted scores.

One of the most comprehensive evaluations in this section was conducted by Rupp et al. (2019), who used e-rater® with 9,628 English essays written by Swiss and German high-school students. Four essay prompts were used from the Test of English as a Foreign Language (TOEFL) website; two prompts required students to "state, explain, and support an opinion" (p. 2), and the other two prompts required students to read and listen to a lecture and then "critically relate the information in the two sources" (p. 2). Teams of human raters were trained to reach a high level of agreement on a 6-point scale; *QWK* values for human-human agreement ranged from .62 to .87. Next, 96 classifier models were trained with 50% of the data using five machine learning techniques. The human-machine agreement was better for prompt-specific models than generic models. Furthermore, generic models trained with the study's data performed better than "off-the-shelf" generic models that were pre-trained with a large sample of previous TOEFL responses. The two best-performing models were prompt-specific SVM models trained with pooled data, with *QWK* values ranging from .72 to .81. The authors conclude that the custom-built prompt-specific models "performed generally satisfactorily" (p. 16) but that "automated scoring generally pays off only at large operational volumes and remains, at least at the moment, the purview of assessment and learning companies that intend to provide at-scale solutions." (p. 18).

Other articles used Criterion®, a web-based system that incorporates e-rater® to score essays and Critique to provide feedback (e.g., Aluthman, 2016). Ma and Slater (2015) compared how essays written by EFL college students were scored by Criterion® versus instructors and third-party graders. Additionally, they conducted a focus group wherein three writing instructors reviewed and discussed six of the graded essays. Pearson correlation coefficients were low between the system's scores and instructor grades ($r = .39$). However, the correlations were higher between the system's scores and third-party graders (.61) than between the instructors and third-party graders (.35). Interpretation is limited as the study did not report accuracies or F1 scores. Nevertheless, the focus group provided insights on how instructors approached grading. For example, grammatical errors tended to be forgiven more than word choice and use of pronouns "since inappropriate word choice prevented the teachers from understanding the intended meaning, and unclear pronouns had the teachers wondering what they referred to" (p. 411). This qualitative data suggests that different errors were weighted differently by human graders depending on how the errors impede understanding.

McNamara et al. (2015) introduced a hierarchical system of categorizing features to reflect more accurately how humans grade essays. They hypothesized that "expert raters might engage in something similar to a sorting task, initially grouping essays based on relatively superficial criteria, and subsequently classifying essays based on finer-grained characteristics of the essay" (p. 39). Thus, they developed a system that first partitioned essays by overall length before assigning scores to subsets using different sets of predictive features. These features were indexed using three instruments; from a total of 440 variables, 320 variables correlated with human scores, and 140 remained after exclusion for multicollinearity. When this hierarchical approach was applied to a corpus of 1243 argumentative essays, written in response to 14 essay prompts and by three grade levels, overall exact and adjacent accuracies were 55% and 92%, respectively.

Coh-Metrix, one of the instruments used in McNamara et al. (2015), deserves further description as some versions are freely available online. Coh-Metrix consists of 106 features organized into nine broad categories (Latifi & Gierl, 2021), ranging from descriptive indices (e.g., number of words, sentences, and paragraphs) to referential cohesion and lexical diversity[1]. Other articles also used Coh-Metrix, although with differing levels of complexity. For example, Petchprasert (2021) simply used six components from Coh-Metrix to compare English essays written at the start and end of an 8-week study. In contrast, Perin and Lauterbach (2018) used the complete set of features to predict whether persuasive essays and summaries written by community college students would be classed as high or low proficiency. Despite extensive training, the human-human agreement was particularly poor for the summaries (exact agreement = 23%, vs. 77% for essays). After texts were categorized as either high or low proficiency using a median split, a discriminant analysis correctly classified 66% of persuasive essays and 68% of summaries. However, low F1 scores in the essays showed generally poor human-machine agreement (.25-.78). Additionally, across 104 variables, significant differences were reported across proficiency groups for only two variables for the essays and eight variables for the summaries. As the authors acknowledge, "given the large number of variables tested, these may be chance findings" (p. 72).

Latifi and Gierl (2021) demonstrate the utility of open-access resources like Coh-Metrix. Their study used Coh-Metrix to extract features and Weka[2], another open-source software, to score essays from the Automated Student Assessment Prize (ASAP) competition dataset[3]. This freely available dataset contains 12,978 pre-scored essays written by 7th to 10th grade students on eight essay prompts. Another study in the present corpus (Beseiso et al., 2021) used the dataset to test the performance of a pre-trained transformer-based neural network (RoBERTa), and the frequent use of this dataset in AES research highlights its value (e.g., Kumar et al., 2020; Mayfield & Black,

---

[1] http://cohmetrix.com/

[2] https://www.cs.waikato.ac.nz/ml/weka

[3] https://www.kaggle.com/competitions/asap-aes/data

2020). These open-source resources lower the barrier to entry for smaller teams of researchers who may not have the resources to collect large sets of expert-scored essays or to develop custom feature-extraction and classification algorithms. However, challenges still remain for developing AES systems in non-English languages.

Three articles evaluated AES systems for non-English essays (Alsanie et al., 2021; García-Gorrostieta et al., 2018; Yang et al., 2019). Yang et al. (2019) highlighted the challenges of developing an AEE system for Chinese text. The differences in grammar and structure limited the use of standard tools developed for English text. Thus, the authors had to develop custom methods for extracting features; for example, developing a custom probabilistic segmentation model for word segmentation on Chinese characters, and gathering multiple text corpora for error detection. These corpora consisted of text extracted from multiple sources (e.g., textbooks and newspapers) and organized into common nouns, common misspellings, similar words, and homophones[4]. While the authors could use a dataset from the Chinese Grammatical Error Diagnosis competition to train a neural network to detect grammatical errors, they commented on the relative unavailability of standard Chinese essay datasets to evaluate the accuracy of automated scoring. After collecting a set of 1000 pre-scored essays written by 3rd grade students, models were trained with 170 essays to predict scores on a 6-point scale. When tested on 30 held-out essays, the best-performing model was a random forest with a grammatical error detection component ($QWK = .76$). This study highlighted the engineering challenges involved in developing tools to address characteristics specific to Chinese text, and the value of having open-source standard datasets for AES research.

Similarly, Alsanie et al. (2021) used custom-built algorithms to develop their AES system for Arabic essays. The authors discussed the complex morpho-syntactic rules in the Arabic language and tested parse-tree structures to represent sentences. A SVM was trained to score the syntax of each sentence, and the scores were averaged for an overall syntax score. A linear regression model then combined the syntax score with other features (a semantic score, the number or percentage of spelling mistakes, and total word count) to predict the overall score of 293 essays on a 4-point scale. The essays were written by learners of Arabic as a second language and marked by two teachers. While the level of agreement between human raters was somewhat low ($QWK \approx .44$, exact agreement $\approx$ 42%), higher agreement was reported between human raters and the best performing regression model ($QWK \approx$ .67-.68, exact agreement $\approx$ 45-51%). Models that incorporated the syntax score performed better than those that did not. Therefore, different languages may place greater importance on different parts of writing, and adding features that reflect these language-specific weights can improve the similarity between automated and human scores.

Developing custom algorithms may further spur cross-pollination and development. For example, Raković et al. (2021) leveraged both open-source packages and custom-built indices to explore the rhetorical and content features that differentiated “knowledge-telling” versus “knowledge-transforming” sentences in argumentative essays. Their random forest classifier achieved 73% accuracy. Relatedly, Hanlon et al. (2021) attempted to identify “reflective moves” using Academic Writing Analytics (AWA), a system developed by the University of Technology Sydney[5]. These studies demonstrate the trend towards measuring ever more complex dimensions of writing.

Evaluations of User Experience

Other articles evaluated AES systems by assessing users’ experiences and attitudes. In a large-scale investigation, Wilson et al. (2021) evaluated a district-wide implementation of MI Write, a successor of PEG. The authors assessed correlations between measures extracted from the system’s log data (e.g., total number of essays submitted, number of revisions per essay, and system use across time) with academic and outcome measures (e.g., pre- and post-test scores, state-test writing scores, and teacher and student attitude surveys). The actual use of the system varied greatly across classrooms, and even within the same classroom across time, as teachers could decide independently when and how to use the system. Nevertheless, both students and teachers reported generally positive attitudes towards using MI Write, and generally agreed the system helped students improve their writing.

With the variance in actual usage, the metrics extracted from log data were generally not significant predictors of external outcomes. The total number of essays submitted was a weak predictor of state test

---

[4] Portions of the data are available at: https://github.com/yangyiqin-tsinghua/Automated-Grader-for-Chinese-Essay

[5] https://cic.uts.edu.au/open-source-writing-analytics/

performance for two out of three grade levels; however, the reported effect sizes (β = .08 and .07 for Grades 3 and 5, respectively) were overshadowed by the large negative effect of "the percentage of students within the school receiving free or reduced-priced lunch" (p. 249; β = -.57, -.39, and -.57 for Grades 3, 4, and 5, respectively). These results suggest that although AWE may contribute to improving learners' writing, they are not a silver bullet to overcome broader society-level factors that affect a learner's ability to thrive.

Other studies also assessed whether automatically-generated formative feedback would improve students' writing performance. Wang (2020) assessed the experience of 178 undergraduate students in an EFL course in China. During the course, students wrote four essays which they uploaded onto three AEE systems (Pigai, iWrite, and Awrite), which provided individualized feedback on elements such as grammar, mechanics, and syntactic complexity. Students generally rated the automated feedback positively in surveys, and follow-up interviews suggested that students were satisfied with the feedback on content but expected more feedback on discourse elements. Students scored higher on the last essay than on the first essay. However, the single-group, repeated-measures design of the study limits conclusions about the effects of automated feedback on writing performance. Students may have shown similar levels of improvement even without using AEEs, and any effects of AEEs cannot be separated from the effects of possible confounding variables (e.g., time and practice).

A similar critique can be made regarding other studies that used pre-post designs. For example, Kostikova and Miasoiedova (2019) reported improvements in the writing ability of 32 Ukrainian postgraduate students at the end versus the start of a course that integrated Write & Improve. Similarly, Aluthman (2016) reported higher scores on the tenth than the first essay written by undergraduate EFL students in Saudi Arabia. However, as with Wang (2020), it is uncertain whether any observed improvement can be attributed to the effects of automated feedback.

In contrast, Chodorow et al. (2010) analysed data previously collected in a between-groups design conducted by Lipnevich and Smith (2009). In Lipnevich and Smith's study, 464 undergraduate students wrote a 500-word essay and were then separated into three groups: A control group received no feedback and the other two groups received feedback generated by Criterion®. One of the experimental groups was told the feedback was computer-generated, and the other that the feedback was from the instructor. After encouragement to reread and revise their essays, students uploaded a final version to Criterion®. Chodorow et al. assessed the number of article errors in a subset of the essays and reported significant main effects of language (native vs. non-native) and version (first vs revised essay) but not feedback (p = .06). After relative error rates were split into three categories (better, worse, and no-change), the authors reported a larger proportion of non-native speakers in the feedback groups improved in the revised essay (.53), compared with the no-feedback control (.36). While this categorical comparison cannot show how much the computer-generated feedback contributed to decreased error rates, the results of the between-groups design suggest that the feedback did somewhat contribute.

In the sole study that evaluated both the performance of a system and its users' experiences, Knight et al. (2018) presented a preliminary mixed-method assessment of AWA. The system was used to identify markers of eight different rhetoric moves (e.g., summarising an issue, contrasting ideas, emphasising ideas, and providing background knowledge). The system employed pattern-matching between the input essay and lexicons of words and expressions associated with each rhetoric move (e.g., the pattern "challenge, need, failure and shift" was associated with the move "contrast"; p. 5). The AWA system was trialled by 40 volunteers who submitted 3000-word argumentative essays on a topic in civil law. The AWA tool highlighted sentences that it recognised as a rhetoric move and bolded the words that triggered the classification. A law lecturer reviewed a small subset of the highlighted sentences to fill the cells of a confusion matrix. While this annotation exercise did not produce enough counts to calculate reliable F1 scores, the confusion matrix showed high levels of false alarms that were caused by standard phrases in law (e.g., "discovery" and "dispute resolution") which were mistaken as markers of rhetorical moves. Consequently, removing these words from the lexicons improved classification.

The AWA system was generally well-received, with multiple respondents noting that the system was a useful source of feedback for reflecting on their writing. However, some also raised concerns like the system not detecting some forms of "summarising language" and the possibility of the system constricting their writing style. Conclusions from this qualitative data are limited as only 12/40 students completed the 4-item survey, possibly indicating self-selection bias in the sample of responses. Nevertheless, this study was a laudable attempt to assess both the accuracy of the system and the experience of its users using both quantitative and qualitative methods.

Conclusion

The articles in this corpus give a snapshot of the AES research literature, but should not be mistaken for a comprehensive overview of all AES research. Our review considered only primary research articles published in peer-reviewed journals; this leaves out book chapters that review and discuss AES (e.g., Mayfield & Rose, 2013) and research published as conference proceedings or in online repositories (e.g., Kumar et al., 2020; Mayfield & Black, 2020). Furthermore, inclusion/exclusion judgments involved grey areas such as the distinction between essays and short answers. We included articles that explicitly used automatic scoring on “essays” and excluded those on short-answer texts (e.g., Gierl et al., 2014; Maestrales et al., 2021; Noyes et al., 2020) and answers requiring less than three paragraphs (e.g., Nehm et al., 2012; Passonneau et al., 2018). Nevertheless, these articles used similar feature-extraction and classification techniques, and the short length of some “essays” (e.g., in the ASAP dataset) blurs this distinction further. Therefore, we mention some of these omissions for interested readers. Future systematic reviews focusing solely on AES could cast a wider net by including conference proceedings and explicitly integrating the plethora of AES-related terms into their search string.

Despite its limitations, the present synthesis reveals interesting trends in AES research. Common research designs are generally limited to evaluating either a system’s accuracy or its user’s experiences, with only one article in the present corpus that evaluated both (Knight et al., 2018). Furthermore, single-group pre-post designs limit the conclusions that can be drawn. In contrast, studies using mixed-method designs demonstrate the value of combining qualitative and quantitative data to compare the processes of human versus machine scoring (e.g., Ma & Slater, 2015). In line with international policy work on ethical AI (e.g., Hu et al, 2019; IEEE, 2019; UNESCO, 2021), we recommend that future AES research continues to pursue interdisciplinary research between computer scientists and educators.

References

Alsanie, W., Alkanhal, M. I., Alhamadi, M., & Alqabbany, A. O. (2021). Automatic scoring of Arabic essays over three linguistic levels. *Progress in Artificial Intelligence*, *11*, 1–13. https://doi.org/10.1007/s13748-021-00257-z

Aluthman, E. S. (2016). The effect of using automated essay evaluation on ESL undergraduate students' writing skill. *International Journal of English Linguistics*, *6*, 54. https://doi.org/10.5539/ijel.v6n5p54

Balfour, S. P. (2013). Assessing writing in MOOCs: Automated essay scoring and Calibrated Peer Review™. *Research & Practice in Assessment*, *8*, 40-49.

Beseiso, M., Alzubi, O. A., & Rashaideh, H. (2021). A novel automated essay scoring approach for reliable higher educational assessments. *Journal of Computing in Higher Education*, *33*, 727–746. https://doi.org/10.1007/s12528-021-09283-1

Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing*, *27*, 419–436. https://doi.org/10.1177/0265532210364391

Dikli, S. (2010). The nature of automated essay scoring feedback. *CALICO Journal*, *28*, 99–134. https://doi.org/10.11139/cj.28.1.99-134

Ericcson, P. F., & Haswell, R. H. (Eds.). (2006). *Machine scoring of student essays: Truth and consequences*. Utah State University Press. https://digitalcommons.usu.edu/usupress_pubs/139

Foltz, P.W. Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, & Computers*, *28*, 197–202. https://doi.org/10.3758/BF03204765

García-Gorrostieta, J. M., López-López, A., & González-López, S. (2018). Automatic argument assessment of final project reports of computer engineering students. *Computer Applications in Engineering Education*, *26*, 1217–1226. https://doi.org/10.1002/cae.21996

Gierl, M. J., Latifi, S., Lai, H., Boulais, A.-P., & De Champlain, A. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical Education*, *48*, 950–962. https://doi.org/10.1111/medu.12517

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, *36*, 193-202. https://doi.org/10.3758/BF03195564

Gough, D., Oliver, S., & Thomas, J. (2017). *An introduction to systematic reviews* (2nd ed.). Sage.

Hanlon, C. D., Frosch, E. M., Shochet, R. B., Buckingham Shum, S. J., Gibson, A., & Goldberg, H. R. (2021). Recognizing reflection: Computer-assisted analysis of first year medical students' reflective writing. *Medical Science Educator*, *31*, 109–116. https://doi.org/10.1007/s40670-020-01132-7

Hu, X., Neupane, B., Echaiz, L. F., Sibal, P., & Rivera Lam, M. (2019). Steering AI and advanced ICTs for knowledge societies: A Rights, Openness, Access, and Multi-stakeholder Perspective. UNESCO Publishing. https://unesdoc.unesco.org/ark:/48223/pf0000372132

IEEE. (2019). Ethically *Aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems* (1st Ed.). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html

Ifenthaler, D. (in press). Automated essay scoring systems. In O. Zawacki-Richter, & I. Jung (Eds.), *Handbook of Open, Distance and Digital Education*. Springer.

Knight, S., Buckingham Shum, S., Ryan, P., Sándor, Á., & Wang, X. (2018). Designing Academic Writing Analytics for civil law student self-assessment. *International Journal of Artificial Intelligence in Education*, *28*, 1–28. https://doi.org/10.1007/s40593-016-0121-0

Kostikova, I. I., & Miasoiedova, S. V. (2019). Supporting post-graduate students writing skills development with the online machine learning tool: Write & Improve. *Information Technologies and Learning Tools*, *74*, 238–249. https://doi.org/10.33407/itlt.v74i6.2600

Kumar, Y., Bhatia, M., Kabra, A., Li, J. J., Jin, D., & Shah, R. R. (2020). Calling out bluff: attacking the robustness of automatic scoring systems with simple adversarial testing. *arXiv preprint arXiv:2007.06796*. https://onikle.com/articles/293757

Latifi, S., & Gierl, M. (2021). Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing. *Language Testing*, *38*, 62–85. https://doi.org/10.1177/0265532220929918

Lipnevich, A. A., & Smith, J. K. (2009). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied*, *15*, 319-333. https://doi.org/10.1037/a0017841

Lu, X. (2019). An Empirical Study on the artificial intelligence writing evaluation system in China CET. *Big Data*, *7*, 121–129. https://doi.org/10.1089/big.2018.0151

Ma, H., & Slater, T. (2015). Using the developmental path of cause to bridge the gap between AWE scores and writing teachers' evaluations. *Writing & Pedagogy*, *7*, 395–422. https://doi.org/10.1558/wap.v7i2-3.26376

Mayfield, E., & Black, A. W. (2020). Should you fine-tune BERT for Automated Essay Scoring?. In J. Burstein, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis, & T. Zesch (Eds.). *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 151-162). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.bea-1.15

Mayfield, E., & Rosé, C. P. (2013). LightSIDE Open source machine learning for text. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 124-135). Taylor & Francis.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23*, 35–59. https://doi.org/10.1016/j.asw.2014.09.002

Nazari, N., Shabbir, M. S., & Setiawan, R. (2021). Application of artificial intelligence powered digital writing assistant in higher education: Randomized controlled trial. *Heliyon*, *7*(5), e07014. https://doi.org/10.1016/j.heliyon.2021.e07014

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, *21*, 183-196. https://doi.org/10.1007/s10956-011-9300-9

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, *47*, 238–243. http://www.jstor.org/stable/20371545

Passonneau, R. J., Poddar, A., Gite, G., Krivokapic, A., Yang, Q., & Perin, D. (2018). Wise crowd content assessment and educational rubrics. *International Journal of Artificial Intelligence in Education*, *28*, 29–55. https://doi.org/10.1007/s40593-016-0128-6

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). Parlor Press.

Perelman, L. (2020). The BABEL generator and E-rater: 21st century writing constructs and automated essay scoring (AES). *Journal of Writing Assessment*, *13*. https://escholarship.org/uc/item/263565cq

Perin, D., & Lauterbach, M. (2018). Assessing text-based writing of low-skilled college students. *International Journal of Artificial Intelligence in Education*, *28*, 56–78. https://doi.org/10.1007/s40593-016-0122-z

Petchprasert, A. (2021). Utilizing an automated tool analysis to evaluate EFL students' writing performances. *Asian-Pacific Journal of Second and Foreign Language Education*, *6*, 1. https://doi.org/10.1186/s40862-020-00107-w

Raković, M., Winne, P. H., Marzouk, Z., & Chang, D. (2021). Automatic identification of knowledge-transforming content in argument essays developed from multiple sources. *Journal of Computer Assisted Learning*, *37*, 903–924. https://doi.org/10.1111/jcal.12531

Rupp, A. A., Casabianca, J. M., Krüger, M., Keller, S., & Köller, O. (2019). Automated essay scoring at scale: A case study in Switzerland and Germany. *ETS Research Report Series*, *2019*, 1–23. https://doi.org/10.1002/ets2.12249

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000381137

Wang, Z. (2020). Computer-assisted EFL writing and evaluations based on artificial intelligence: A case from a college reading and writing course. *Library Hi Tech*, *40*, 80–97. https://doi.org/10.1108/LHT-05-2020-0113

Wilson, J., Huang, Y., Palermo, C., Beard, G., & MacArthur, C. A. (2021). Automated feedback and automated scoring in the elementary grades: Usage, attitudes, and associations with writing outcomes in a districtwide implementation of MI Write. *International Journal of Artificial Intelligence in Education*, *31*, 234–276. https://doi.org/10.1007/s40593-020-00236-w

Yang, Y., Xia, L., & Zhao, Q. (2019). An automated grader for Chinese essay combining shallow and deep semantic attributes. *IEEE Access*, *7*, 176306–176316. https://doi.org/10.1109/ACCESS.2019.2957582

Zawacki-Richter, O., Kerres, M., Bedenlier, S., Bond, M., & Buntins, K. (Eds.). (2020). *Systematic reviews in educational research: Methodology, perspectives and application*. Springer. https://doi.org/10.1007/978-3-658-27602-7

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education*, *16*, 1-27. https://doi.org/10.1186/s41239-019-0171-0