Automated Essay Scoring Systems

Dirk Ifenthaler

Contents

Abstract

Essays are scholarly compositions with a specific focus on a phenomenon in question. They provide learners the opportunity to demonstrate in-depth understanding of a subject matter; however, evaluating, grading, and providing feedback on written essays are time consuming and labor intensive. Advances in automated assessment systems may facilitate the feasibility, objectivity, reliability, and validity of the evaluation of written prose as well as providing instant feedback during learning processes. Measurements of written text include observable components such as content, style, organization, and mechanics. As a result, automated essay scoring systems generate a single score or detailed evaluation of predefined assessment features. This chapter describes the evolution and features of automated scoring systems, discusses their limitations, and concludes with future directions for research and practice.

D. Ifenthaler (✉)
University of Mannheim, Mannheim, Germany

Curtin University, Perth, WA, Australia
e-mail: dirk@ifenthaler.info

O. Zawacki-Richter, I. Jung (eds.), *Handbook of Open, Distance and Digital Education*,
https://doi.org/10.1007/978-981-19-0351-9_59-1

Keywords

Automated essay scoring · Essay grading system · Writing assessment · Natural language processing · Educational measurement · Technology-enhanced assessment · Automated writing evaluation

Introduction

Educational assessment is a systematic method of gathering information or artifacts about a learner and learning processes to draw inferences of the persons' dispositions (E. Baker, Chung, & Cai, 2016). Various forms of assessments exist, including single- and multiple-choice, selection/association, hot spot, knowledge mapping, or visual identification. However, using natural language (e.g., written prose or essays) is regarded as the most useful and valid technique for assessing higher-order learning processes and learning outcomes (Flower & Hayes, 1981). Essays are scholarly analytical or interpretative compositions with a specific focus on a phenomenon in question. Valenti, Neri, and Cucchiarelli (2003) as well as Zupanc and Bosnic (2015) note that written essays provide learners the opportunity to demonstrate higher order thinking skills and in-depth understanding of a subject matter. However, evaluating, grading, and providing feedback on written essays are time consuming, labor intensive, and possibly biased by an unfair human rater.

For more than 50 years, the concept of developing and implementing computer-based systems, which may support automated assessment and feedback of written prose, has been discussed (Page, 1966). Technology-enhanced assessment systems enriched standard or paper-based assessment approaches, some of which hold much promise for supporting learning processes and learning outcomes (Webb, Gibson, & Forkosh-Baruch, 2013; Webb & Ifenthaler, 2018). While much effort in institutional and national systems is focused on harnessing the power of technology-enhanced assessment approaches in order to reduce costs and increase efficiency (Bennett, 2015), a range of different technology-enhanced assessment scenarios have been the focus of educational research and development, however, often at small scale (Stödberg, 2012). For example, technology-enhanced assessments may involve a pedagogical agent for providing feedback during a learning process (Johnson & Lester, 2016). Other scenarios of technology-enhanced assessments include analyses of a learners' decisions and interactions during game-based learning (Bellotti, Kapralos, Lee, Moreno-Ger, & Berta, 2013; Kim & Ifenthaler, 2019), scaffolding for dynamic task selection including related feedback (Corbalan, Kester, & van Merriënboer, 2009), remote asynchronous expert feedback on collaborative problem-solving tasks (Rissanen et al., 2008), or semantic rich and personalized feedback as well as adaptive prompts for reflection through data-driven assessments (Ifenthaler & Greiff, 2021; Schumacher & Ifenthaler, 2021).

It is expected that such technology-enhanced assessment systems meet a number of specific requirements, such as (a) adaptability to different subject domains, (b) flexibility for experimental as well as learning and teaching settings,

(c) management of huge amounts of data, (d) rapid analysis of complex and unstructured data, (e) immediate feedback for learners and educators, as well as (f) generation of automated reports of results for educational decision-making.

Given the on-going developments in computer technology, data analytics, and artificial intelligence, there are advances in automated assessment systems, which may facilitate the feasibility, objectivity, reliability, and validity of the assessment of written prose as well as providing instant feedback during learning processes (Whitelock & Bektik, 2018). Accordingly, automated essay grading (AEG) systems, or automated essay scoring (AES systems, are defined as a computer-based process of applying standardized measurements on open-ended or constructed-response text-based test items. Measurements of written text include observable components such as content, style, organization, mechanics, and so forth (Shermis, Burstein, Higgins, & Zechner, 2010). As a result, the AES system generates a single score or detailed evaluation of predefined assessment features (Ifenthaler, 2016).

This chapter describes the evolution and features of automated scoring systems, discusses their limitations, and concludes with future directions for research and practice.

Synopsis of Automated Scoring Systems

The first widely known automated scoring system, Project Essay Grader (PEG), was conceptualized by Ellis Battan Page in late 1960s (Page, 1966, 1968). PEG relies on proxy measures, such as average word length, essay length, number of certain punctuation marks, and so forth, to determine the quality of an open-ended response item. Despite the promising findings from research on PEG, acceptance and use of the system remained limited (Ajay, Tillett, & Page, 1973; Page, 1968). The advent of the Internet in the 1990s and related advances in hard- and software introduced a further interest in designing and implementing AES systems. The developers primarily aimed to address concerns with time, cost, reliability, and generalizability regarding the assessment of writing. AES systems have been used as a co-rater in large-scale standardized writing assessments since the late 1990s (e.g., e-rater by Educational Testing Service). While initial systems focused on English language, a wide variety of languages have been included in further developments, such as Arabic (Azmi, Al-Jouie, & Hussain, 2019), Bahasa Malay (Vantage Learning, 2002), Hebrew (Vantage Learning, 2001), German (Pirnay-Dummer & Ifenthaler, 2011), or Japanese (Kawate-Mierzejewska, 2003). More recent developments of AES systems utilize advanced machine learning approaches and elaborated natural language processing algorithms (Glavas, Ganesh, & Somasundaran, 2021).

For almost 60 years, different terms related to automated assessment of written prose have been used mostly interchangeably. Most frequently used terms are automated essay scoring (AES) and automated essay grading (AEG); however, more recent research used the term automated writing evaluation (AWE) and automated essay evaluation (AEE) (Zupanc & Bosnic, 2015). While the above-mentioned system focuses on written prose including several hundred words,

another field developed focusing on short answers referred to as automatic short answer grading (ASAG) (Burrows, Gurevych, & Stein, 2015).

Functions of Automated Scoring Systems

AES systems mimic human evaluation of written prose by using various methods of scoring, that is, statistics, machine learning, and natural language processing (NLP) techniques. Implemented features of AES systems vary widely, yet they are mostly trained with large sets of expert-rated sample open-ended assessment items to internalize features that are relevant to human scoring. AES systems compare the features in training sets to those in new test items to find similarities between high/low scoring training and high/low scoring new ones and then apply scoring information gained from training sets to new item responses (Ifenthaler, 2016).

The underlying methodology of AES systems varies; however, recent research mainly focuses on natural language processing approaches (Glavas et al., 2021). AES systems focusing on content use Latent Semantic Analysis (LSA) which assumes that terms or words with similar meaning occur in similar parts of written text (Wild, 2016). Other content-related approaches include Pattern Matching Techniques (PMT). The idea of depicting semantic structures, which include concepts and relations between the concepts, has its source in two fields: semantics (especially propositional logic) and linguistics. Semantic oriented approaches include Ontologies and Semantic Networks (Pirnay-Dummer, Ifenthaler, & Seel, 2012). A semantic network represents information in terms of a collection of objects (nodes) and binary associations (directed labeled edges), the former standing for individuals (or concepts of some sort), and the latter standing for binary relations over these. Accordingly, a representation of knowledge in a written text by means of a semantic network corresponds with a graphical representation where each node denotes an object or concept, and each labeled being one of the relations used in the knowledge representation. Despite the differences between semantic networks, three types of edges are usually contained in all network representation schemas (Pirnay-Dummer et al., 2012): (a) Generalization: connects a concept with a more general one. The generalization relation between concepts is a partial order and organizes concepts into a hierarchy. (b) Individualization: connects an individual (token) with its generic type. (c) Aggregation: connects an object with its attributes (parts, functions) (e.g., wings – part of – bird). Another method of organizing semantic networks is partitioning which involves grouping objects and elements or relations into partitions that are organized hierarchically, so that if partition A is below partition B, everything visible or present in B is also visible in A unless otherwise specified (Hartley & Barnden, 1997).

From an information systems perspective, understood as a set of interrelated components that accumulate, process, store, and distribute information to support decision making, several preconditions and processes are required for a functioning AES system (Burrows et al., 2015; Pirnay-Dummer & Ifenthaler, 2010):

1. Assessment scenario: The assessment task with a specific focus on written prose needs to be designed and implemented. Written text is being collected from learners and from experts (being used as a reference for later evaluation).
2. Preparation: The written text may contain characters which could disturb the evaluation process. Thus, a specific character set is expected. All other characters may be deleted. Tags may be also deleted, as are other expected metadata within each text.
3. Tokenizing: The prepared text gets split into sentences and tokens. Tokens are words, punctuation marks, quotation marks, and so on. Tokenizing is somewhat language dependent, which means that different tokenizing methods are required for different languages.
4. Tagging: There are different approaches and heuristics for tagging sentences and tokens. A combination of rule-based and corpus-based tagging seems most feasible when the subject domain of the content is unknown to the AES system. Tagging and the rules for it is a quite complex field of linguistic methods (Brill, 1995).
5. Stemming: Specific assessment attributes may require that flexions of a word will be treated as one (e.g., the singular and plural forms "door" and "doors"). Stemming reduces all words to their word stems.
6. Analytics: Using further natural language processing (NLP) approaches, the prepared text is analyzed regarding predefined assessment attributes (see below), resulting in models and statistics.
7. Prediction: Further algorithms produce scores or other output variables based on the analytics results.
8. Veracity: Based on available historical data or reference data, the analytics scores are compared in order to build trust and validity in the AES result.

Common assessment attributes of AES have been identified by Zupanc and Bosnic (2017) including linguistic (lexical, grammar, mechanics), style, and content attributes. Among 28 lexical attributes, frequencies of characters, words, sentences are commonly used. More advanced lexical attributes include average sentence length, use of stopwords, variation in sentence length, or the variation of specific words. Other lexical attributes focus on readability or lexical diversity utilizing specific measures such as Gunning Fox index, Nominal ratio, Type-token-ratio (DuBay, 2007). Another 37 grammar attributes are frequently implemented, such as number of grammar errors, complexity of sentence tree structure, use of prepositions and forms of adjectives, adverbs, nouns, verbs. A few attributes focus on mechanics, for example, the number of spellchecking errors, the number of capitalization errors, or punctuation errors. Attributes that focus on content include similarities with source or reference texts or content-related patterns (Attali, 2011). Specific semantic attributes have been described as concept matching and proposition matching (Ifenthaler, 2014). Both attributes are based on similarity measures (Tversky, 1977). Concept matching compares the sets of concepts (single words) within a written text to determine the use of terms. This measure is especially important for different assessments which operate in the same domain. Propositional

matching compares only fully identical propositions between two knowledge representations. It is a good measure for quantifying complex semantic relations in a specific subject domain. Balanced semantic matching measure uses both concepts and propositions to match the semantic potential between the knowledge representations. Such content or semantic oriented attributes focus on the correctness of content and its meaning (Ifenthaler, 2014).

Overview of Automated Scoring Systems

Instructional applications of automated scoring systems are developed to facilitate the process of scoring and feedback in writing classrooms. These AES systems mimic human scoring by using various attributes; however, implemented attributes vary widely.

The market of commercial and open-source AES systems has seen a steady growth since the introduction of PEG. The majority of available AES systems extract a set of attributes from written prose and analyze it using some algorithm to generate a final output. Several overviews document the distinct features of AES systems (Dikli, 2011; Ifenthaler, 2016; Ifenthaler & Dikli, 2015; Zupanc & Bosnic, 2017). Burrows et al. (2015) identified five eras throughout the almost 60 years of research in AES: (1) concept mapping, (2) information extraction, (3) corpus-based methods, (4) machine learning, and (5) evaluation.

Zupanc and Bosnic (2017) note that four commercial AES systems have been predominant in application: PEG, e-rater, IEA, and IntelliMetric. Open access or open code systems have been available for research purposes (e.g., AKOVIA); however, they are yet to be made available to the general public. Table 1 provides an overview of current AES systems, including a short description of the applied assessment methodology, output features, information about test quality, and specific requirements. The overview is far from being complete; however, it includes major systems which have been reported in previous summaries and systematic literature reviews on AES systems (Burrows et al., 2015; Dikli, 2011; Ifenthaler, 2016; Ifenthaler & Dikli, 2015; Ramesh & Sanampudi, 2021; Zupanc & Bosnic, 2017). Several AES systems also have instructional versions for classroom use. In addition to their instant scoring capacity on a holistic scale, the instructional AES systems are capable of generating diagnostic feedback and scoring on an analytic scale as well. The majority of AES systems use focus on style or content-quality and use NLP algorithms in combination with variations of regression models. Depending on the methodology, AES system requires training samples for building a reference for future comparisons. However, the test quality, precision, or accuracy of several AES systems is publicly not available or has not been reported in rigorous empirical research (Wilson & Rodrigues, 2020).

Table 1 Overview of AES systems

AES system	Methodology	Output	Quality	Requirements
CRASE	Statistics and NLP; machine learning; style and content-quality	Score on an essay, short constructed response item, and graphic item	N/A	75 responses for training samples and 500 responses for cross-validation
IEA	LSA, NLP, machine learning, content quality	Score, customizable dashboard	Reliability and validity studies	100–300 training samples
e-rater	NLP, linear regression, style and content quality	Holistic and analytic score; immediate feedback on traits through its instructional application (Criterion)	Reliability and validity studies	465 training samples
Benchmark-SkillWriter	NLP, neural networks, style and content quality	Analytic scores, rubric scales, and immediate feedback	Reliability and validity studies	N/A
IntelliMetric	NLP, statistical model, style and content quality	Holistic and analytic score, immediate feedback on traits through its instructional application (MY Access)	Reliability and validity studies	300 training samples
AKOVIA	NLP, statistical model, similarity matching, structure and content quality	Customizable feedback including immediate score, written and graphical feedback	Reliability and validity studies	None, requires reference text/ model
PEG	Statistical model, style	Holistic and analytic scoring; immediate feedback on traits through its instructional application (PEG Writing)	Reliability and validity studies	100–400 training samples
Markit	NLP, pattern matching, linear regression, content quality	Score on an essay	N/A	1 reference essay
LightSIDE	Machine learning, multilevel modeling techniques; content-quality	Score on an essay	N/A	300 training samples

(continued)

Table 1 (continued)

AES system	Methodology	Output	Quality	Requirements
Lexile	NLP, Lexile measure, style and content quality	Score on text characteristics	N/A	0
SAGrader	Fuzzy logic, rule-based analysis, semantics	Score on semantics, immediate feedback through its instructional application	N/A	0
BETSY	Bayesian text classification, style and content quality	Trait scoring and feedback	Reliability and validity studies	1000 training samples

Note. NLP = Natural Language Processing; LSA = Latent Semantic Analysis

Open Questions and Directions for Research

There are several concerns regarding the precision of AES systems and the lack of semantic interpretation capabilities of underlying algorithms. Reliability and validity of AES systems have been extensively investigated (Landauer, Laham, & Foltz, 2003; Shermis et al., 2010). The correlations and agreement rates between AES systems and expert human raters have been found to be fairly high; however, the agreement rate is not at the desired level yet (Gierl, Latifi, Lai, Boulais, & Champlain, 2014). It should be noted that many of these studies highlight the results of adjacent agreement between humans and AES systems rather than those of exact agreement (Ifenthaler & Dikli, 2015). Exact agreement is harder to achieve as it requires two or more raters to assign the same exact score on an essay while adjacent agreement requires two or more raters to assign a score within one scale point of each other. It should also be noted that correlation studies are mostly conducted at high-stakes assessment settings rather than classroom settings; therefore, AES versus human inter-rater reliability rates may not be the same in specific assessment settings. The rate is expected to be lower in the latter since the content of an essay is likely to be more important in low-stakes assessment contexts.

The validity of AES systems has been critically reflected since the introduction of the initial applications (Page, 1966). A common approach for testing validity is the comparison of scores from AES systems with those of human experts (Attali & Burstein, 2006). Accordingly, questions arise about the role of AES systems promoting purposeful writing or authentic open-ended assessment responses, because the underlying algorithms view writing as a formulaic act and allows writers to concentrate more on the formal aspects of language such as origin, vocabulary, grammar, and text length with little or no attention to the meaning of the text (Ifenthaler, 2016). Validation of AES systems may include the correct use of specific assessment attributes, the openness of algorithms, and underlying aggregation and

analytics techniques, as well as a combination of human and automated approaches before communicating results to learners (Attali, 2013). Closely related to the issue of validity is the concern regarding reliability of AES systems. In this context, reliability assumes that AES systems produce repeatedly consistent scores within and across different assessment conditions (Zupanc & Bosnic, 2015). Another concern is the bias of underlying algorithms, that is, algorithms have their source in a human programmer which may introduce additional error structures or even features of discrimination (e.g., cultural bias based on selective text corpora). Criticism has been put toward commercial marketing of AES systems for speakers of English as a second or foreign language (ESL/EFL) when the underlying methodology has been developed based on English language with native-English speakers in mind. In an effort to assist ESL/EFL speakers in writing classrooms, many developers have incorporated a multilingual feedback function in the instructional versions of AES systems. Receiving feedback in the first language has proven benefits, yet it may not be sufficient for ESL/EFL speakers to improve their writing in English. It would be more beneficial for non-native speakers of English if developers take common ESL/EFL errors into consideration when they build algorithms in AES systems. Another area of concern is that writers can trick AES systems. For instance, if the written text produced is long and includes certain type of vocabulary that the AES system is familiar with, an essay can receive a higher score from AES regardless of the quality of its content. Therefore, developers have been trying to prevent cheating by users through incorporating additional validity algorithms (e.g., flagging written text with unusual elements for human scoring) (Ifenthaler & Dikli, 2015). The validity and reliability concerns result in speculations regarding the credibility of AES systems considering that the majority of the research on AES is conducted or sponsored by the developing companies. Hence, there is a need for more research that addresses the validity and reliability issues raised above and preferably those conducted by independent researchers (Kumar & Boulanger, 2020).

Despite the above-mentioned concerns and limitation, educational organizations choose to incorporate instructional applications of AES systems in classrooms, mainly to increase student motivation toward writing and reducing workload of involved teachers. They assume that if AES systems assist students with the grammatical errors in their writings, teachers will have more time to focus on content related issues. Still, research on students' perception on AES systems and the effect on motivation as well as on learning processes and learning outcomes is scarce (Stephen, Gierl, & King, 2021). In contrast, educational organizations are hesitant in implementing AES systems mainly because of validity issues related to domain knowledge-based evaluation. As Ramesh and Sanampudi (2021) exemplify, the domain-specific meaning of "cell" may be different in biology or physics. Other concerns that may lower the willingness to adopt of AES systems in educational organizations include fairness, consistency, transparency, privacy, security, and ethical issues (Ramineni & Williamson, 2013; Shermis, 2010).

AES systems can make the result of an assessment available instantly and may produce immediate feedback whenever the learner needs it. Such instant feedback

provides autonomy to the learner during the learning process, that is, learners are not depended on possibly delayed feedback from teachers. Several attributes implemented in AES systems can produce an automated score, for instance, correctness of syntactic aspects. Still, the automated and informative feedback regarding content and semantics is limited. Alternative feedback mechanisms have been suggested, for example, Automated Knowledge Visualization and Assessment (AKOVIA) provides automated graphical feedback models, generated on the fly, which have been successfully tested for preflection and reflection in problem-based writing tasks (Lehmann, Haehnlein, & Ifenthaler, 2014). Other studies using AKOVIA feedback models highlight the benefits of availability of informative feedback whenever the learner needs it and its identical impact on problem solving when compared with feedback models created by domain experts (Ifenthaler, 2014).

Questions for future research focusing on AES systems may focus on (a) construct validity (i.e., comparing AES systems with other systems or human rater results), (b) interindividual and intraindividual consistency and robustness of AES scores obtained (e.g., in comparison with different assessment tasks), (c) correlative nature of AES scores with other pedagogical or psychological measures (e.g., interest, intelligence, prior knowledge), (d) fairness and transparency of AES systems and related scores, as well as (e) ethical concerns related to AES systems, (f) (Elliot & Williamson, 2013). From a technological perspective, (f) the feasibility of the automated scoring system (including training of AES using prescored, expert/reference, comparison) is still a key issue with regard to the quality of assessment results. Other requirements include the (g) instant availability, accuracy, and confidence of the automated assessment. From a pedagogical perspective, (h) the form of the open-ended or constructed-response test needs to be considered. The (i) assessment capabilities of the AES system, such as the assessment of different languages, content-oriented assessment, coherence assessment (e.g., writing style, syntax, spelling), domain-specific features assessment, and plagiarism detection, are critical for a large-scale implementation. Further, (j) the form of feedback generated by the automated scoring system might include simple scoring but also rich semantic and graphical feedback. Finally, (k) the integration of an AES system into existing applications, such as learning management systems, needs to be further investigated by developers, researchers, and practitioners.

Implications for Open, Distance, and Digital Education

The evolution of Massive Open Online Courses (MOOCs) nurtured important questions about online education and its automated assessment (Blackmon & Major, 2017; White, 2014). Education providers such as Coursera, edX, and Udacity dominantly apply so-called auto-graded assessments (e.g., single- or multiple-choice assessments). Implementing automated scoring for open-ended assessments is still on the agenda of such provides, however, not fully developed yet (Corbeil, Khan, & Corbeil, 2018).

With the increased availability of vast and highly varied amounts of data from learners, teachers, learning environments, and administrative systems within educational settings, further opportunities arise for advancing AES systems in open, distance, and digital education. Analytics-enhanced assessment enlarges standard methods of AES systems through harnessing formative as well as summative data from learners and their contexts in order to facilitate learning processes in near real-time and help decision-makers to improve learning environments. Hence, analytics-enhanced assessment may provide multiple benefits for students, schools, and involved stakeholders. However, as noted by Ellis (2013), analytics currently fail to make full use of educational data for assessment.

Interest in collecting and mining large sets of educational data on student background and performance has grown over the past years and is generally referred to as learning analytics (R. S. Baker & Siemens, 2015). In recent years, the incorporation of learning analytics into educational practices and research has further developed. However, while new applications and approaches have brought forth new insights, there is still a shortage of research addressing the effectiveness and consequences with regard to AES systems. Learning analytics, which refers to the use of static and dynamic data from learners and their contexts for (1) the understanding of learning and the discovery of traces of learning and (2) the support of learning processes and educational decision-making (Ifenthaler, 2015), offers a range of opportunities for formative and summative assessment of written text. Hence, the primary goal of learning analytics is to better meet students' needs by offering individual learning paths, adaptive assessments and recommendations, or adaptive and just-in-time feedback (Gašević, Dawson, & Siemens, 2015; McLoughlin & Lee, 2010), ideally, tailored to learners' motivational states, individual characteristics, and learning goals (Schumacher & Ifenthaler, 2018). From an assessment perspective focusing on AES systems, learning analytics for formative assessment focuses on the generation and interpretation of evidence about learner performance by teachers, learners, and/or technology to make assisted decisions about the next steps in learning and instruction (Ifenthaler, Greiff, & Gibson, 2018; Spector et al., 2016). In this context, real- or near-time data are extremely valuable because of their benefits in ongoing learning interactions. Learning analytics for written text from a summative assessment perspective is utilized to make judgments that are typically based on standards or benchmarks (Black & Wiliam, 1998).

In conclusion, analytics-enhanced assessments of written essays may reveal personal information and insights into an individual learning history; however, they are not accredited and far from being unbiased, comprehensive, and fully valid at this point in time. Much remains to be done to mitigate these shortcomings in a way that learners will truly benefit from AES systems.

Cross-References

► Artificial Intelligence in Education and Ethics
► Design, Delivery and Assessment: Introduction and Conclusions

▶ Feedback and Assessment in Online Learning
▶ Learner Support System
▶ Learning Analytics
▶ Self- and Peer-Assessment in ODDE
▶ Technology Applications and LMS

References

Ajay, H. B., Tillett, P. I., & Page, E. B. (1973). *The analysis of essays by computer (AEC-II). Final report*. Storrs, CT: University of Connecticut.

Attali, Y. (2011). *A differential word use measure for content analysis in automated essay scoring*. ETS Research Report Series, 36.

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181–198). New York, NY: Routledge.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V. 2. *The Journal of Technology, Learning and Assessment, 4*(3), 3–29. https://doi.org/10.1002/j.2333-8504.2004.tb01972.x.

Azmi, A., Al-Jouie, M. F., & Hussain, M. (2019). AAEE – Automated evaluation of students' essays in Arabic language. *Information Processing & Management, 56*(5), 1736–1752. https://doi.org/10.1016/j.ipm.2019.05.008.

Baker, E., Chung, G., & Cai, L. (2016). Assessment, gaze, refraction, and blur: The course of achievement testing in the past 100 years. *Review of Research in Education, 40*, 94–142. https://doi.org/10.3102/0091732X16679806.

Baker, R. S., & Siemens, G. (2015). Educational data mining and learning analytics. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 253–272). Cambridge, UK: Cambridge University Press.

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013). Assessment in and of serious games: An overview. *Advances in Human-Computer Interaction, 2013*, 136864. https://doi.org/10.1155/2013/136864.

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education, 39*(1), 370–407. https://doi.org/10.3102/0091732x14554179.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. https://doi.org/10.1080/0969595980050102.

Blackmon, S. J., & Major, C. H. (2017). Wherefore art thou MOOC?: Defining massive open online courses. *Online Learning Journal, 21*(4), 195–221. https://doi.org/10.24059/olj.v21i4.1272.

Brill, E. (1995). Unsupervised learning of dismabiguation rules for part of speech tagging. Paper presented at the Second Workshop on Very Large Corpora, WVLC-95, Boston. Paper presentation retrieved from

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education, 25*(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8.

Corbalan, G., Kester, L., & van Merriënboer, J. J. G. (2009). Dynamic task selection: Effects of feedback and learner control on efficiency and motivation. *Learning and Instruction, 19*(6), 455–465. https://doi.org/10.1016/j.learninstruc.2008.07.002.

Corbeil, J. R., Khan, B. H., & Corbeil, M. E. (2018). MOOCs revisited: Still transformative or passing fad? *Asian Journal of University Education, 14*(2), 1–12.

Dikli, S. (2011). The nature of automated essay scoring feedback. *CALICO Journal, 28*(1), 99–134. https://doi.org/10.11139/cj.28.1.99-134.

DuBay, W. H. (2007). *Smart language: Readers, readability, and the grading of text*. Costa Mesa, CA, USA: BookSurge Publishing.

Elliot, N., & Williamson, D. M. (2013). Assessing writing special issue: Assessing writing with automated scoring systems. *Assessing Writing, 18*(1), 1–6. https://doi.org/10.1016/j.asw.2012.11.002.

Ellis, C. (2013). Broadening the scope and increasing usefulness of learning analytics: The case for assessment analytics. *British Journal of Educational Technology, 44*(4), 662–664. https://doi.org/10.1111/bjet.12028.

Flower, L., & Hayes, J. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365–387.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends, 59*(1), 64–71. https://doi.org/10.1007/s11528-014-0822-x.

Gierl, M. J., Latifi, S., Lai, H., Boulais, A.-P., & Champlain, A. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical Education, 48*(10), 950–962. https://doi.org/10.1111/medu.12517.

Glavas, G., Ganesh, A., & Somasundaran, S. (2021). Training and domain adaptation for supervised text segmentation. Paper presented at the Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, Virtual Conference.

Hartley, R. T., & Barnden, J. A. (1997). Semantic networks: Visualizations of knowledge. *Trends in Cognitive Science, 1*(5), 169–175. https://doi.org/10.1016/S1364-6613(97)01057-7.

Ifenthaler, D. (2014). AKOVIA: Automated knowledge visualization and assessment. *Technology, Knowledge and Learning, 19*(1–2), 241–248. https://doi.org/10.1007/s10758-014-9224-6.

Ifenthaler, D. (2015). Learning analytics. In J. M. Spector (Ed.), *The SAGE encyclopedia of educational technology* (Vol. 2, pp. 447–451). Thousand Oaks, CA: Sage.

Ifenthaler, D. (2016). Automated grading. In S. Danver (Ed.), *The SAGE encyclopedia of online education* (p. 130). Thousand Oaks, CA: Sage.

Ifenthaler, D., & Dikli, S. (2015). Automated scoring of essays. In J. M. Spector (Ed.), *The SAGE encyclopedia of educational technology* (Vol. 1, pp. 64–68). Thousand Oaks, CA: Sage.

Ifenthaler, D., & Greiff, S. (2021). Leveraging learning analytics for assessment and feedback. In J. Liebowitz (Ed.), *Online learning analytics* (pp. 1–18). Boca Raton, FL: Auerbach Publications.

Ifenthaler, D., Greiff, S., & Gibson, D. C. (2018). Making use of data for assessments: Harnessing analytics and data science. In J. Voogt, G. Knezek, R. Christensen, & K.-W. Lai (Eds.), *International handbook of IT in primary and secondary education* (2nd ed., pp. 649–663). New York, NY: Springer.

Johnson, W. L., & Lester, J. C. (2016). Face-to-face interaction with pedagogical agents, twenty years later. *International Journal of Artificial Intelligence in Education, 26*(1), 25–36. https://doi.org/10.1007/s40593-015-0065-9.

Kawate-Mierzejewska, M. (2003). *E-rater software*. Paper presented at the Japanese Association for Language Teaching, Tokyo, Japan. Paper presentation retrieved from

Kim, Y. J., & Ifenthaler, D. (2019). Game-based assessment: The past ten years and moving forward. In D. Ifenthaler & Y. J. Kim (Eds.), *Game-based assessment revisted* (pp. 3–12). Cham, Switzerland: Springer.

Kumar, V. S., & Boulanger, D. (2020). Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-020-00211-5.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the intelligent essay assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Erlbaum.

Lehmann, T., Haehnlein, I., & Ifenthaler, D. (2014). Cognitive, metacognitive and motivational perspectives on preflection in self-regulated online learning. *Computers in Human Behavior, 32*, 313–323. https://doi.org/10.1016/j.chb.2013.07.051.

McLoughlin, C., & Lee, M. J. W. (2010). Personalized and self regulated learning in the Web 2.0 era: International exemplars of innovative pedagogy using social software. *Australasian Journal of Educational Technology, 26*(1), 28–43.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47*(5), 238–243.

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education, 14*(2), 210–225. https://doi.org/10.1007/BF01419938.

Pirnay-Dummer, P., & Ifenthaler, D. (2010). Automated knowledge visualization and assessment. In D. Ifenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 77–115). New York, NY: Springer.

Pirnay-Dummer, P., & Ifenthaler, D. (2011). Text-guided automated self assessment. A graph-based approach to help learners with ongoing writing. In D. Ifenthaler, K. P. Isaias, D. G. Sampson, & J. M. Spector (Eds.), *Multiple perspectives on problem solving and learning in the digital age* (pp. 217–225). New York, NY: Springer.

Pirnay-Dummer, P., Ifenthaler, D., & Seel, N. M. (2012). Semantic networks. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (Vol. 19, pp. 3025–3029). New York, NY: Springer.

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

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing, 18*(1), 25–39. https://doi.org/10.1016/j.asw.2012.10.004.

Rissanen, M. J., Kume, N., Kuroda, Y., Kuroda, T., Yoshimura, K., & Yoshihara, H. (2008). Asynchronous teaching of psychomotor skills through VR annotations: Evaluation in digital rectal examination. *Studies in Health Technology and Informatics, 132*, 411–416.

Schumacher, C., & Ifenthaler, D. (2018). The importance of students' motivational dispositions for designing learning analytics. *Journal of Computing in Higher Education, 30*(3), 599–619. https://doi.org/10.1007/s12528-018-9188-y.

Schumacher, C., & Ifenthaler, D. (2021). Investigating prompts for supporting students' self-regulation – A remaining challenge for learning analytics approaches? *The Internet and Higher Education, 49*, 100791. https://doi.org/10.1016/j.iheduc.2020.100791.

Shermis, M. D. (2010). Automated essay scoring in a high stakes testing environment. In V. J. Shute & B. J. Becker (Eds.), *Innovative assessment for the 21st century* (pp. 167–184). New York, NY: Springer.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In P. Petersen, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (pp. 75–80). Oxford, England: Elsevier.

Spector, J. M., Ifenthaler, D., Sampson, D. G., Yang, L., Mukama, E., Warusavitarana, A., … Gibson, D. C. (2016). Technology enhanced formative assessment for 21st century learning. *Educational Technology & Society, 19*(3), 58–71.

Stephen, T. C., Gierl, M. C., & King, S. (2021). Automated essay scoring (AES) of constructed responses in nursing examinations: An evaluation. *Nurse Education in Practice, 54*, 103085. https://doi.org/10.1016/j.nepr.2021.103085.

Stödberg, U. (2012). A research review of e-assessment. *Assessment & Evaluation in Higher Education, 37*(5), 591–604. https://doi.org/10.1080/02602938.2011.557496.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327–352.

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education, 2*, 319–330.

Vantage Learning. (2001). *A preliminary study of the efficacy of IntelliMetric ® for use in scoring Hebrew assessments*. Retrieved from Newtown, PA:

Vantage Learning. (2002). *A study of IntelliMetric ® scoring for responses written in Bahasa Malay (No. RB-735)*. Retrieved from Newtown, PA:

Webb, M., Gibson, D. C., & Forkosh-Baruch, A. (2013). Challenges for information technology supporting educational assessment. *Journal of Computer Assisted Learning, 29*(5), 451–462. https://doi.org/10.1111/jcal.12033.

Webb, M., & Ifenthaler, D. (2018). Section introduction: Using information technology for assessment: Issues and opportunities. In J. Voogt, G. Knezek, R. Christensen, & K.-W. Lai (Eds.),

International handbook of IT in primary and secondary education (2nd ed., pp. 577–580). Cham, Switzerland: Springer.

White, B. (2014). Is "MOOC-mania" over? In S. S. Cheung, J. Fong, J. Zhang, R. Kwan, & L. Kwok (Eds.), *Hybrid learning. Theory and practice* (Vol. 8595, pp. 11–15). Cham, Switzerland: Springer International Publishing.

Whitelock, D., & Bektik, D. (2018). Progress and challenges for automated scoring and feedback systems for large-scale assessments. In J. Voogt, G. Knezek, R. Christensen, & K.-W. Lai (Eds.), *International handbook of IT in primary and secondary education* (2nd ed., pp. 617–634). New York, NY: Springer.

Wild, F. (2016). *Learning analytics in R with SNA, LSA, and MPIA*. Heidelberg, Germany: Springer.

Wilson, J., & Rodrigues, J. (2020). Classification accuracy and efficiency of writing screening using automated essay scoring. *Journal of School Psychology, 82*, 123–140. https://doi.org/10.1016/j.jsp.2020.08.008.

Zupanc, K., & Bosnic, Z. (2015). Advances in the field of automated essay evaluation. *Informatica, 39*(4), 383–395.

Zupanc, K., & Bosnic, Z. (2017). Automated essay evaluation with semantic analysis. *Knowledge-Based Systems, 120*, 118–132. https://doi.org/10.1016/j.knosys.2017.01.006.