Received: 27 January 2020 | Revised: 21 April 2020 | Accepted: 6 June 2021

DOI: 10.1111/jcal.12577

REVIEW ARTICLE

Journal of Computer Assisted Learning WILEY

Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?'

John Gardner[1] | Michael O'Leary[2] | Li Yuan[3]

[1]Faculty of Social Sciences, University of Stirling, Stirling, UK

[2]Centre for Assessment Research, Policy and Practice in Education (CARPE), Institute of Education, Dublin City University, Dublin, Ireland

[3]College of Education for the Future, Beijing Normal University, Zhuhai, Guangdong, China

Correspondence
Michael O'Leary, Centre for Assessment Research, Policy and Practice in Education (CARPE), Dublin City University, Dublin, Ireland.
Email: michael.oleary@dcu.ie

Abstract

Artificial Intelligence is at the heart of modern society with computers now capable of making process decisions in many spheres of human activity. In education, there has been intensive growth in systems that make formal and informal learning an anytime, anywhere activity for billions of people through online open educational resources and massive online open courses. Moreover, new developments in Artificial Intelligence-related educational assessment are attracting increasing interest as means of improving assessment efficacy and validity, with much attention focusing on the analysis of the large volumes of process data being captured from digital assessment contexts. In evaluating the state of play of Artificial Intelligence in formative and summative educational assessment, this paper offers a critical perspective on the two core applications: automated essay scoring systems and computerized adaptive tests, along with the Big Data analysis approaches to machine learning that underpin them.

KEYWORDS

artificial intelligence, automated essay scoring, big data, computerized adaptive tests, learning analytics, machine learning

1 | ARTIFICIAL INTELLIGENCE AND EDUCATIONAL ASSESSMENT

The second part of this paper's title might seem strange to some but others might recognize the words used by the editor of Phi Delta Kappan to describe a horizon-scanning piece by Ellis Page in 1966. Somewhat incredulous, the editor's preamble was trying, but failing, to keep an open mind on Page's proposition that sometime, in the then near future, computers would relieve English teachers of the burden of essay marking. Page proclaimed that the time was coming when computers would give teachers 'a stylistic and subject-matter analysis … and extensive comment and suggestion' on their students' work 'by the first bell the next day' (Page, 1966, p. 239). Page was in effect proposing that automated essay scoring (AES) of writing would soon be on a par with that of human assessors. As time has passed, AES has indeed become established as a very sophisticated tool for technical aspects of essay writing in large-scale summative testing programmes and there are rapidly developing formative feedback applications for using the AES process to assist learners in improving their writing. However, some controversy surrounds AES use whenever it is suggested that it assesses the quality of writing; as it is somewhat less amenable to assessing the creative and higher-order dimensions of writing. In parallel with AESs, and often combined with them, is another powerful genre of AI-related assessment, namely computerised adaptive testing (CAT). This is primarily a summative tool but also has potential for framing its outcomes as purposeful formative feedback to learners.

There is intense interest in AI applications for education (e.g., see Tuomi et al., 2018; UNESCO, 2019) and as this grows, increasing interest is being shown in applications for educational assessment. The essence of artificial intelligence (AI) in both summative and

formative contexts is the concept of machine 'learning' – where the computer is 'taught' how to interpret patterns in data and 'trained' to undertake predetermined actions according to those interpretations. This machine intelligence has arguably facilitated many of the huge step-changes underpinning the transformation to the 21st century's information society – in all areas including commerce, manufacturing, health and the relatively new phenomenon of social media. In fact, it continues to impact upon almost every aspect of life in countries and communities where economic affluence enables people to exploit its many applications – with major economic blocs investing heavily in AI applications and talent development (see Castro et al., 2019). Using today's massive computational power, large data sets –Big Data – are captured from online processes involved in every aspect of modern living (e.g., technology, medicine, environment, commerce) and are subjected to a variety of essentially correlational and probabilistic analyses to identify patterns of prior behaviour and to predict or propose next actions. The computer learns the strength of the associations in these Big Data sets and, whether its next action is to propose an aerodynamic refinement for an aircraft wing, remand a crime suspect in custody to protect the public (see Partnership on AI, n.d.), link two 'lonely hearts' together or predict the likelihood of a typhoon event in the Philippines, it continues to learn from new data.

Paraphrasing McKinsey and Co (2011) and Gartner Glossary (2019), Big Data is broadly conceived as any very large-scale and dynamically growing collection of 'information assets', which require a level of intelligent computer-based analysis that is beyond the capacity of ordinary data-processing hardware and software. Gartner's definition characterizes the features of this dynamic growth of information assets as the three Vs: huge *Volumes* of *Variable* types of data being processed at varying speeds (*Velocity*). Big Data is a concept that has its origins in the massive computing contexts of science. Billions of readings from experiments are subjected to large-scale algorithmic analysis in pursuit of patterns, causation and predictions in fields as diverse as engineering, quantum physics and astronomy. Big Data analysis has also become a staple aspect of such major dimensions of modern society as medical diagnosis, consumer trend analysis and weather forecasting.

Arguably, the 'intelligent' characteristics of these applications are developed at two levels: unsupervised and supervised. In the case of the former, that is the untrained or unsupervised machine, the computer simply identifies patterns in massive data sets for subsequent interpretation by human experts. In the latter, the machine is trained (i.e., supervised) by human experts both to identify and learn specific patterns in the massive data sets and to effect automated actions in relation to them. Clearly, the former can often be a preceding step in setting up the processes of the latter.

In educational assessment, the same underlying concepts of machine learning apply. If the computer can be 'taught' the content that students are required to know and can ask questions to which it has 'learned' the answers, it can assess those students on their knowledge. In a more sophisticated step, if the computer can learn what quality criteria apply to a student's understanding and application of that knowledge in relation to an assessment task, whether written or verbal (e.g., see Somasundaran et al., 2015), and can learn how to identify these criteria in the student's responses, it has the potential to assess the quality of the work. The distinction between having knowledge and being able to understand and apply it will not be lost on educators, as it is along this continuum that the capabilities of human judges and machine assessments ultimately part company. This is particularly evident in the computer-based assessment of student essays.

2 | AUTOMATED ESSAY SCORING

Page (1966) alluded to two types of elements in student essays: the *trin* (an intrinsic property of the student work such as fluency) and a *prox* (such as the length of the essay as a proxy variable related to fluency, or the number of commas as a proxy 'measure' of good punctuation). He foresaw a time in the future when natural language processing (NLP) would achieve the technical maturity to enable machines to learn and understand how to assess the existence of the many complex trins in human writing. However, back in 1966, when he and his team were manually setting up students' essays on punch cards, he was quite reasonably settling for small steps and was targeting a relatively small set of 31 'proxes' or proxies as the mainstay of his Project Essay Grading system. Fast-forward forty years and Ben-Simon and Bennett (2007) were identifying the four leading, commercial AES systems as: Project Essay Grading (from the original Page work), Intellimetric (developed by Vantage Technologies), Intelligent Essay Assessor (initially University of Colorado and more recently Pearson Education) and e-Rater (ETS). It could be argued that remarkably little has changed in the fundamentals of automated essay assessment since Page's work, for example in the types of data used. However, the AI engines that drive AESs today have changed greatly in the sophistication of their algorithms, data capacity and processing efficiency; and as a result their details are kept secret for commercial reasons. Nevertheless, research activities and reports over time have highlighted the main elements of writing, which these engines commonly seek to detect.

Among the organizations promoting their AESs, the US assessment giant, ETS, has a long history of making public their research on the e-Rater system, through engagement with the academic and user communities. For example, Deane (2013) confirmed that ETS's e-Rater (v. 11.1) uses four main elements of writing, with relatively easily detected proxies:

- Grammar (e.g., incorrect subject-verb agreement, incorrect pronouns, possessive errors)
- Usage (e.g., article and preposition errors, incorrect word forms)
- Mechanics (e.g., errors in capitalization, punctuation, spellings)
- Style (e.g., repetitious word use)

and a selection of elements with more complex proxies such as:

- Discourse structure (e.g., presence of a thesis statement, main points, length of discourse elements)
- Lexical complexity (e.g., use of unusual/sophisticated words)
- Sentence variety

- Source use
- Discourse coherence quality (e.g., extent of off-topic writing)

One example of a freely available system, TAALES (Kyle et al., 2018), is a specialized tool for identifying levels of lexical sophistication (derived originally from Graesser et al.'s, 2004 Coh-Metrix) that tracks over 400 different indices including word frequency, word range, academic language, word recognition norms and contextual distinctiveness. A feature of these non-commercial systems is the openness of related publications and Crossley and Kyle (2018), for example, offer candid and interesting insights into the workings of TAALES and, by extension, AESs more generally. Most AESs use varieties of the types of proxies above though how they use them tends to be a closed secret.

To arrive at a score, AES systems must first learn how these proxies relate to quality in the learners' responses, and in order to do this the computer must be trained on human raters' assessments of a range of student work. In simple terms, the instances of proxy variables occurring are identified in a large sample of candidates' scripts and compared with the scores given by a panel of judges. The more judges and scripts there are, the more acceptable is the final average rating and, most importantly, the weighting of each proxy's contribution to that rating across the scripts. Using appropriate weightings, seemingly remarkable agreement levels of up to 80% have been reported between essay rating systems and the assessments that human judges give the same scripts. Although not always the case, many of the claims for high correlation can be attributed to how the comparisons are made, for example summative scores for the essays being on an ordinal scale of 1–6, or sometimes 1–4. In these simple correlational contexts, it might be reasonable to expect a high level of agreement at the extremes of the rating range; for example, very bad and very good essays may reasonably be expected to exhibit correspondingly more or less of the indicator proxies. In contrast, variance is most likely to occur in the essays that are tricky to assess; for example, where higher order aspects of the writing (creativity, analysis, argument, synthesis etc., which the computer fails to detect) are privileged and rewarded by human judges over the perhaps weaker – but machine detectable – mechanics, grammar or usage features.

In the beginning (e.g., with Page and his contemporaries) the driving motivation was to make the large-scale assessment of essays, efficient, consistent and low-cost through the use of AESs. The well-known variation in human raters' scores for even the same essay (for reasons that the computer may not easily learn, e.g., differences of opinion on what constitutes 'originality'), require moderation processes for mitigation in the non-AES world. A significant number of judges is always needed to try to even out the anomalies that will inevitably arise between human judges – and at national scale, this can be a huge cost. Therefore, for several decades, commercial AES systems have sought to establish credible parity of performance with human judges.

However, for the community of teachers of English in the Anglophone and second language worlds, the technical elements of writing enjoy an arguably secondary role to a much more complex and comprehensive conception of writing as a construct. The US Framework for Success in Postsecondary Writing (CWPA, NCTE, & NWP, 2011) signals the complexity of this construct in which teachers of writing seek to prepare their students' readiness for higher education by developing such skills as the use of:

- Rhetorical knowledge – the ability to analyse and act on understandings of audiences, purposes, and contexts in creating and comprehending texts;
- Critical thinking – the ability to analyse a situation or text and make thoughtful decisions based on that analysis, through writing, reading and research;
- Writing processes – multiple strategies to approach and undertake writing and research;
- Knowledge of conventions – the formal and informal guidelines that define what is considered to be correct and appropriate, or incorrect and inappropriate, in a piece of writing;
- Ability to compose in multiple environments – from traditional pen and paper to electronic technologies.

The Framework further describes eight 'habits of mind', considered essential for success in college writing:

- Curiosity – the desire to know more about the world.
- Openness – the willingness to consider new ways of being and thinking in the world.
- Engagement – a sense of investment and involvement in learning.
- Creativity – the ability to use novel approaches for generating, investigating, and representing ideas.
- Persistence – the ability to sustain interest in and attention to short- and long-term projects.
- Responsibility – the ability to take ownership of one's actions and understand the consequences of those actions for oneself and others.
- Flexibility – the ability to adapt to situations, expectations or demands.
- Metacognition – the ability to reflect on one's own thinking as well as on the individual and cultural processes used to structure knowledge

The challenges for AES developers have therefore added to the original triad of goals, namely efficient, consistent and low-cost assessment of writing tasks, especially in large-scale testing contexts such as MOOCs, to a much more specific pursuit of sophisticated proxies that might advance the cause of the valid assessment of writing as a comprehensive, complex construct. NLP is the primary development tool in these endeavours and is considerably bolstered by second-language learning research. For example, ETS's resource pages offer a number of papers on its innovative uses in tracking students' usage of metaphor (Beigman Klebanov et al., 2015), opinion (Farra et al., 2015) and sources (Beigman Klebanov et al., 2014), and on other complex features such as grammaticality (Heilman et al., 2014). Whilst this latter paper on grammaticality may be excellent in its context, it highlights one of the key features of AES systems that attract the criticism of writing experts, criticisms that have persisted down through the

decades since Page's visionary piece in 1966. The feature in question is essentially the reductionism that is a fundamental part of all AI systems – vast collections of individual, countable units of data – Big Data – are subjected to massive analytic processes that enable the training of the computer to model and learn how different patterns in the data correlate with targeted properties of the data-set. In the grammaticality paper, quality assessments were predicted on a comparable basis to human ratings (on a simple ranked scale of 1–4, with 1 = incomprehensible and 4 = perfect) for samples taken from a large data-set of over 3000 sentences in essays written by non-native speakers of English. Clearly, there is potential in such tools to be part of a more sophisticated AI-NLP approach to essay assessment but the very fact that the assessment of an important aspect of writing could be reduced to a single sentence-level analysis and short ordinal scale is something akin to a red rag to a bull for many writing experts. Nevertheless, with the crock of commercial testing gold awaiting the designer of the more-human–than-human AES, development efforts have been in full swing in the last decade or so (e.g., Hewlett Packard, 2012).

Opposition to AES approaches to assessment of writing quality comes in a number of forms. For example, the US National Council of Teachers of English has delivered a withering critique of machine scoring, informed by a sizeable list of automated assessment-related publications (NCTE, 2013). One of the most well-known academic critics is Perelman whose writings claim that automated scoring of essays is simply a nonsense (though he often uses more colourful language to make his point, e.g., Perelman, 2012a). He has analysed and even 'gamed' (Perelman, 2018) a number of systems to provide evidence for his case against AESs. For example, he derides e-Rater as rewarding 'the use of jargon and obscure and pretentious language' (Perelman, 2012b, p 126) and scathingly dismisses all AESs on the charge 'that they do not understand meaning and they are not sentient' (ibid, p 125). He also accuses them of consistently over-privileging essay length in their assessments (Perelman, 2014). Some AES providers do little to endear themselves to the legion of critics who take the same position as Perelman. For example, Pearson's promotional literature on the Intelligent Essay Assessor claims that it can '… "understand" the meaning of text much the same as a human reader' (Pearson, n.d.).

Other writing specialists, for example Condon (2013), see the main problem as being the types of test that in his view underrepresent the construct of writing whilst purporting to 'measure' it (e.g., the prevailing definition of an essay is between 300 and 600 words in most AES evaluations). Condon accepts the importance of these general criticisms, and to some extent the use of gaming to highlight AES weaknesses. However, he argues that they are essentially 'red herrings' as they distract from what he feels is the central issue: 'The focus on whether scores rendered by AES systems agree with human raters' scores fails to answer the question of whether these two measurements that are supposed to be related are in fact unrelated' (p. 101). However, he concedes that the AES community does understand and accept the general criticisms, for example citing Deane's (2013, p. 12) acknowledgement that AES systems focus on 'text quality measured in the end product' whilst human raters focus on the student's writing skill. He argues that human assessors read students' texts with the intention of understanding whilst AES systems are designed only to recognize patterns in the texts. Even when the human and computer outcome scores are similar, their inherent meaning is not: 'No AES system can achieve the kind of understanding necessary to evaluate writing on the semantic level – on the level of meaning, let alone the level of awareness of occasion, purpose and audience demanded by any form of real-world writing' (p. 102). He therefore argues strongly against the use of AESs for assessing writing in such summative assessment, high stakes contexts as admissions, placement or achievement testing. Despite this, developments such as those offered by Grammarly.com and Proofreader.com are finding a niche in using AES-type analysis of drafts to help students improve their work before final submission.

Research and development continues to refine automated assessment of writing and has begun to examine how it can support formative assessment feedback to help learners improve their writing. Less useful feedback is the single digit score (0–6 for example) in relation to technical aspects of the writing but when it is augmented with reports on errors and related writing features, which teachers may address and students may reflect upon, it may prove to be a very important formative assessment tool in promoting improved learning. One recent example is the multi-lingual ReaderBench (Botarleanu et al., 2018), which is claimed to track proxies originally identified in a number of systems such as e-Rater and Coh-Metrix. This system is specifically designed to offer formative feedback to learners and teachers using a combination of tools for assessing textual complexity, cohesion, semantics and dialogue features (the latter contributing to feedback on aspects of collaborative and group learning). Nevertheless, the time when AES systems will be able to operate on a par with human judges, with similar levels of connoisseurship for such features as meaning, emotion, originality, creativity, fluency, sense of audience and so on, arguably remains a long way off. Until then the charges will continue to be levelled against AESs: that they do not understand writing and they privilege the reductive technicalities of text. The human assessment of the quality of writing, rendered as overall scores, may therefore continue to be found to correlate with the computer assessments based on these technical features, but the feedback that arises from each type will almost always demonstrate that the AES and human assessments are not assessing the same thing.

3 | COMPUTERIZED ADAPTIVE TESTS

Computerized adaptive tests, CATs, are another core form of machine learned assessment, usually in summative assessment contexts such as high-stakes selection processes (for university entry, employment etc.). By their nature, and as outlined below, this summative orientation predominates and it is only in recent times that applications have begun to be developed for CAT-based processes to be used for formative feedback purposes.

The two best known and arguably most successful CAT-based tests in educational assessment are the Graduate Management

Admissions Council's GMAT (www.mba.com/exams/gmat) and ETS's GRE General examinations (https://www.ets.org/gre). Their formats below give an indication of the role that CAT applications play alongside non-CAT elements (e.g., AESs) in these composite assessment regimes. GMAT results are used by business schools around the globe, for example to select applicants for MBA programmes. The four-part examination comprises two human and/or computer-assessed tests: Integrated Reasoning (12 multiple-response items) and Analytical Writing Assessment (human and AES assessed 'essay'), and two CATs: Verbal Reasoning (36) single answer multiple choice questions (MCQs)) and Quantitative Reasoning (31 MCQs). 'Review and revise' options for completed items or sections are not facilitated.

The GRE General test has six sections: Analytical Writing, Verbal Reasoning (VR) $\times$ 2, Quantitative Reasoning (QR) $\times$ 2) and an unscored section for calibration or research. The VR and QR sections are 'section-level adaptive' meaning that the underlying CAT for these sections determines the difficulty level of the second of each type of section. The Analytical Writing section comprises two 30-minute tasks on 'Analyze an Issue' and 'Analyze an Argument' respectively, both assessed on a human and an AES basis. Both types of section allow examinees to mark items for review within the section time limits.

These two systems follow similar design criteria and have a relatively simple modus operandi (e.g., see Davey, 2011). However, they also have technically complex, algorithmic engines that carry out the item selection and assessment processes. The first distinction of note between CATs and the AESs above is that whilst machine learning in AESs seeks to mimic the judgements of a human rater, a CAT uses a set of test items to position an examinee on a scale that is pre-calibrated for two associated measures: examinees' abilities and item characteristics. CATs use a process in which the *response* an examinee makes to a test *item* enables the computer to purposefully select the next item to assess whether the examinee has yet reached the limit of their ability in the trait under examination.

The underlying psychometry in CATs is based on Item Response Theory, IRT, which first came to the fore when Lord and Novick (1968) and Lord (1970) described a form of its application to 'tailored' tests (aka 'adaptive' tests). Up until this time, Classical Test Theory (CTT) was the most common testing approach offering a test-level assessment of an examinee's ability by using fixed forms of test instruments (the typical pre-defined and static multiple choice 'paper', for example). CTT still has a major role to play in today's world of educational assessment but since Lord and Novick, IRT has underpinned the momentous rise of CAT systems that exploit aspects of machine learning. In contrast to test-level CTT, CATs are item-level tests that dynamically adjust to the examinee's responses to individual items. The adjustments are based on the examinee's demonstrated level of ability in the trait under examination and the characteristics of the items used. With a sufficient item bank size, CAT proponents claim that they are much less vulnerable to the security issues encountered in using fixed tests. They also have lower invigilation demands, may take less time/items to arrive at an acceptable assessment of the examinee's ability and should accurately place the examinee on an ability level that is reproducible in repeat administrations of the test.

The item bank is one of the central features governing success of an IRT based assessment system, and once a suitable set of items has been created and calibrated against a range of pilot examinees, the computer can offer tests, such as non-adaptive linear-on-the-fly-tests (LOFTs) or CATs, that comprise different statistically selected items for each examinee. Each item for a CAT is calibrated on the basis of its item characteristics and on the same scale as examinees' ability levels. For most CATS these characteristics include the item's power to discriminate between examinees of different ability, its level of difficulty (based on the proportion of examinees who answer it correctly in the calibration process) and a third parameter offering an estimate of the probability that its correct answer could be guessed - its 'guessability'. Note that some CATs may use IRT systems that calibrate on only one parameter e.g. difficulty, or two parameters e.g. difficulty and discrimination.

One of the most enduring and best exemplifications of an operational CAT was first published as an online simulation by Rudner (1998). The simple step-by-step description accompanying the simulation explains that:

> 'Computer adaptive testing can begin when an item bank exists with IRT item statistics available on all items, when a procedure has been selected for obtaining ability estimates based upon candidate item performance, and when there is an algorithm chosen for sequencing the set of test items to be administered to candidates.
>
> 'The CAT algorithm is usually an iterative process with the following steps

All the items that have not yet been administered are evaluated to determine which will be the best one to administer next given the currently estimated ability level

1. The 'best' next item is administered and the examinee responds
2. A new ability estimate is computed based on the responses to all of the administered items
3. Steps 1 through 3 are repeated until a stopping criterion is met'

Step 0, so to speak, is the selection of the first item, which prudence suggests should be one that the examinee can be expected to answer correctly. To identify the appropriate difficulty level, the preamble to the test may ask the examinee questions that elicit indicators of their ability (national examination grades etc.) and then a rough estimate of ability can be used to select the first item. In the absence of any 'intelligence' on the examinee's ability, CATs often offer items that the calibration process identifies as being in the lower levels of difficulty (e.g., answered correctly by, say, 70% of test takers). This facility is being increasingly exploited to enable learners to decide, or be guided, on at what point in an online course they should begin

their work. At the other end of the process, the most common stopping criterion is the point when the CAT has decided that the examinee has reached a level of difficulty in the item selections that can reasonably be concluded to be their ceiling. In simpler pass-fail outcome designs, the stopping criterion would be when the examinee surpasses the cut-score or has no prospect of reaching it.

Very few of the fundamentals of CAT design have changed since the time of Lord's specification for 'tailored' tests but technological developments have brought improvements to both previously perceived limitations and to some of the less than ideal features of their application and administration. For example, though the same rules of thumb obtain for item bank size, for example, minimally around 1000 IRT items calibrated against 300+ examinees (Ju & Bork, 2005) to cover the trait under examination, time has allowed some established CATs to develop very substantial item banks (with associated Big Data sets). These banks can be constantly reviewed as each new test administration refines the item characteristics, resulting in more detailed examinee cohort profiles. Related to their importance in servicing CATs, tools for developing item banks and even tools to assess their adaptivity (Reckase et al., 2019) are now forming a significant part of the established infrastructure for CAT development. von Davier (2019) also argues the potential for using deep neural network approaches (DNN) to harvest and transform major static resources (in this case the corpus of medical education texts hosted by PubMed) to offer item writers high quality MCQ item stems and case study material for medical examinations.

For most CATs, item bank development arguably has to be sustained by the fee income from very large candidate volumes. ETS's GRE test, for example, enjoys a huge candidature (416,631 candidates in 2018 in the US and perhaps 200,000 more in over 180 other countries) and at a $205 entry fee the business is on a huge scale (ETS, 2018). The overall cost of running a major examination programme comprises more than just item bank development, of course, but significant sums would be required to keep it refreshed and comprehensive.

CATs work best when dealing with unidimensional content, that is, a relatively well-defined knowledge domain that can be assessed using MCQs with dichotomous outcomes for example, simple correct/incorrect or a correct combination of responses. Some progress is being made on polytomous outcomes, including partially correct responses, but the complexity involved is substantial (e.g., see Aybek & Demirtasli, 2017). When the content domain is less bounded, with several or many subfields each with their own share of the overall knowledge domain, CATs begin to struggle. For example, the notion that numeracy, or specifically arithmetic, is a unidimensional construct or ability would not be readily accepted in educational circles and any CAT will require coverage of items on addition, subtraction, multiplication and division to argue that a measure of ability on arithmetic processes has been achieved. Examinees will differ in their abilities to perform successfully in the sub-domains and at different item characteristic levels (e.g., difficulty levels). It can be difficult to create items to address these types of issues in a balanced manner across the examinee population and Stocking et al. (2000) concluded that even for two sub-domains, large item banks (larger than were available in 2000) would be required for adequate content balancing. One relatively recent solution, therefore, is to use multi-stage test (MST) designs that offer particular sub-domain sections in sequence with appropriate selection algorithms to ensure both item balance and relevant calibration characteristics across the sections. Such staging also presents opportunities for feedback to tutors and learners on content sub-domains in which the learners may struggle or excel. As with all such formative feedback, this can enable tuition to be modified or learners to engage in self-regulated learning (SRL) for improvement.

One feature of effective SRL, the facility that enables an examinee to look over earlier answers and revise them if necessary, is now appearing in certain types of CAT application (for example, in ETS's GRE there is a limited review facility). In the straightforward dichotomous versions, where responses to single correct answer (MCQs are expected to be completed in a timed sequence, such a facility cannot easily be offered because revised answers would obviously disrupt the item selection process if a revision changes the ability level. Today's CATs hold the promise of being able to use ever more sophisticated item selection processes to counter this disruption and allow examinees the facility to review and revise without confounding the process of establishing the ability level.

Very little has changed in the last decade or so in terms of the validity of outcomes from the main CAT applications. Moneta-Koehler et al. (2017), for example, report their poorly performing predictive capacity in relation to candidate's later success in university courses (see also Hall et al., 2017). There are also concerns that their design may contribute to restricting the entry of women and minority groups into key areas such as the sciences. Miller and Stassun (2014), for example, cite ETS in pointing to female candidates scoring on average 80 points less than males on GRE scores, and African Americans scoring 200 points less than white people. Another example is provided by Hauser and Kingsbury (2004) with differential item functioning (DIF) analysis showing up to 25% of the 2003 Idaho grade 10 mathematics test items showing a gender DIF. Interesting examples of CAT developments also show formative potential to address socially contextualized challenges. Wise (2014), for example, has reported on how the problem of item calibration disruption by unmotivated learners may be detected and worked around. This type of research is beginning to push the established summative boundaries of CAT usage with formative processes that that can motivate reluctant examinees.

4 | PROCESS DATA ANALYSIS FOR FORMATIVE ASSESSMENT FEEDBACK

The analysis of large-scale assessment-related datasets is a cornerstone element of AI approaches; for example, for training an AES system or for adaptively deciding the next question presented to a test-taker in a computerized adaptive test. The former is invariably based on many assessors' judgements on aspects of many students' essays; and the latter is based on details of many students' performances on many multiple-choice questions. The unifying feature of the Big Data in these assessment contexts, and the various machine learning

applications in science, medicine and technology, is the concept of Process Data: data that can be purposefully or incidentally captured online as the applications are being used. Techniques for analysing these large volumes of learning and assessment data generally come under the umbrella term, Learning Analytics (sometimes termed Educational Data Mining). The most widely used definition of learning analytics is the one that headlined the first International Conference on Learning Analytics and Knowledge (LAK, 2011; Long & Siemens, 2011): '… the measurement, collection, analysis and reporting of data about learners and their contexts, for the purposes of understanding and optimising learning and the environments in which it occurs'. These are laudable goals but some such as Ellis (2013) argue that a significant proportion of learning analytics activity is pre-occupied with mundane predictions, e.g. identifying patterns in big data sets that are associated with specific outcomes such as improved levels of student achievement.

In relation to assessment, meaningful analysis of the relevant data sets is necessary if the derived information is to be made available to users in an accessible fashion. Users in this context might be those at institutional management level who wish to review institutional-level performance of students, tutors who wish to know whether their teaching has been successful or, indeed, students who wish to know how they have fared within their peer group. Institutional-level analysis of this evaluative nature is more or less de rigueur in today's education settings as institutions seek to analyse 'what works' (or, more to the point, what is not 'working') in their provision. The need to do this will likely be prompted by a combination of formative factors, for example: desiring to facilitate continuous improvements in provision, identifying efficiencies in curriculum delivery, improving the course offerings to prospective students or servicing external accountability requirements. This type of 'academic analytics' is arguably at the low end of intelligent machine analysis and feedback as it is often restricted simply to offering visualization of summative outcomes and trends, for monitoring, marketing or planning purposes.

Cope and Kalantzis (2016) categorize the variety of data to be gathered during a learning process as being machine assessments (e.g. CATs, AESs etc.), structured data that is specifically anticipated and captured by the computer, and unstructured, incidental data, for example DiCerbo and Behrens's 'data exhaust' (DiCerbo & Behrens, 2014). The latter includes the record or 'trace' of time stamps, key strokes and edit histories that clickstream log files can provide. With appropriate analysis, these may point to how a student tackles a problem, the errors and revisions they might make, their misconceptions and even their resilience in the face of making slow or no progress. Furthermore, they argue that dedicated devices such as video cameras, audio recorders, smart watches and bracelets can capture data on eye movements and gaze, facial expressions, body posture, gesture and in-class speech. They can also offer indicators of a wide variety of processes including drafting, peer interaction and even affective states such as confusion, frustration, boredom and level of engagement. Sophisticated analysis of the data captured from intelligent tutor systems (ITSs) can offer tutors insights into how to improve the systems. Molenaar et al. (2019), for example use CAT-type performance assessment data in adaptive learning technologies for selecting appropriate learning resources (instructional materials) or problems for the participating learners to solve. In a similar vein, Lerche and Kiel (2018) have used log data to predict learners' levels of achievement. If the analysis of these various types of data, showing how the learners approach tasks and in what areas they are proficient or are struggling, can be achieved in a timely manner it also has the potential to be mediated to the students as formative feedback. For example, the Embrace system uses trace data dynamically to give young learners immediate formative feedback on their performance in visualized online reading comprehension tasks (Walker et al., 2017).

In another example, Aljohani and Davis (2013) report on how their students are able to use mobile device dashboards to review their quiz results. This gives them immediate feedback on their overall class results, item difficulty-level information on their performance and even a Bloom's taxonomy level of their assessed cognitive understanding. In the manner of the institutional 'academic analytics' above, this is more a descriptive visualization analysis than a machine intelligence (AI) approach but has some merits in providing summative feedback that, in its timeliness, has formative potential.

Thille et al. (2014) argue that the analysis of large-scale assessment data sets can enrich assessment in three main ways: it can be continuous (automatically gathered at all times), feedback-oriented (can be analysed, interpreted and reported for tutors and students) and multi-faceted (can cover the multiplicity of data available through clickstream log files and automated observation). Analysis of large-scale assessment data sets, which continue to grow with each new assessment session and group of test-takers, may also provide a platform for high-level trajectory modelling, which in turn enables individual learners' progress to be compared with typical patterns of progress in the overall student cohort. Importantly, using experts to evaluate student strategies, and to teach the system to give automated 'hints', is a form of scaffolding or formative assessment intervention that such systems can potentially provide at appropriate points in each student's problem solving trajectory. Interestingly, Thille et al also noted that in some instances the strategies proposed by experts, for transitioning from a partial solution to the next stage in solving a problem, ran counter to actual student trajectories and decisions, and the consequent training of the system benefited from a better understanding of how students' approaches varied from experts' expectations. Clearly, this type of learning analytics needs to be very rapid if anything meaningful is to be fed back to students in the live process – raising the potential, according to Cope and Kalantzis (2016) of ending 'the historical separation of instruction and assessment' and for 'feedback that is always available on the fly' (p. 7). Arguably, however, there is limited prospect of these formative assessment techniques migrating any time soon from Thille et al's small-scale on-line environments (ITS, coding practice and MOOC usage) to more diverse learning settings in which data capture is likely to be much more challenging.

Not surprisingly, the growing recognition of the importance of formative assessment in education generally has led to a parallel

interest in using the intelligent analysis of large datasets to assist learners formatively in self-regulating their online learning. Learners' SRL is a complex phenomenon, arguably influenced by a variety of personal traits and circumstantial factors. Cicchinelli et al. (2018), for example, have identified indicators relating to learners' planning and monitoring that associate with higher outcome scores. Others (e.g., Jarvela et al., 2020) argue that the new wave of learning analytics is enabling previously opaque SRL processes to be made visible, even in collaborative learning contexts, through tracing multifaceted affective, social and cognitive indicators. In one such development, the ACT testing group is reporting field-tests on a mobile platform app, Companion, to give students immediate analysis and feedback (ACTNext, 2020). This system uses 'dynamic cognitive diagnostic models and machine learning algorithms' to analyse test results and usage data from a wide selection of learning resources with the promise of full integration into students' daily lives through such vehicles as Amazon's Alexa and Apple's Siri.

Learning analytics feedback for students in any learning context is going to be of maximum usefulness when it takes the form of personalized formative assessment especially in the world of MOOCs or other large-scale e-learning settings, which often have many thousands of contemporaneous learners. In these settings, SRL assumes greater importance because the timeliness or indeed availability of external formative assessment and feedback is seriously constrained by the costs of labour-intensive hot-seating or even asynchronous interaction with tutors. Peer assessment, if accurate (García-Martínez et al., 2019) can help to address this formative deficit but Jansen et al. (2019) have shown that using learning analytics and in-built interventions, that is, in-course video resources on SRL per se, can improve course completion rates in MOOC settings. There seems to be no shortage of proof and near proof of concept in the research literature (see for example, Gutierrez & Crespo Garcia, 2012; Jarvela et al., 2020; Martin & Ndoye, 2016; Tempelaar et al., 2013) but as yet, the holy grail of an off-the-shelf automated and cost effective personalized approach to formative assessment and feedback in MOOCs is top of the wish list for on-line learning developers.

Spector et al. (2016) take the argument for a greater emphasis on personalized formative assessment further and claim that to some extent even ITSs can be one-size-fits-all inasmuch as they identify a learner's specific weakness and provide a remedial response that is the same for all students with the same deficiency. They argue that feedback from learning analytics systems can be dynamically adaptive to the learner through deeper profiling of the learners in combination with the various techniques of performance analysis. Such profiling ranges from the capture of 'stealth' assessments as the student works, described as continuous, embedded and unobtrusive measures of performance designed to identify learner habits, to 'robust' learner profiling including additional data on their preferences, interests and biases. This smacks a little of the on-line profiling of individuals for marketing and other campaigns and may well raise some ethical issues as time goes on – however, as Spector et al concede, such feedback mechanisms 'are yet to be deployed on a large and sustainable scale' (p. 62).

5 | CONCLUDING REMARKS

Our overarching conclusion is that AI in educational assessment has changed little in its basic precepts and functions – that is machine learning and actions based on the results of intelligent analysis of large-scale data – over the last 10 years or so but its technological efficiency, speed and sophistication has advanced on all counts, especially in the analysis of large-scale assessment process data being channelled for formative purposes. Some of the advancement is due to dogged research in universities and research centres but credit must also be given to the large not-for-profit assessment organizations who plough test income into areas of research that simply could not command sufficient funding in the academic world. The core aspects of AI application in this paper, AESs and CATs, have benefited and continue to benefit enormously from technical advances in machine learning. However, the prospect of unilaterally substituting AI judges for human judges in most aspects of student assessment any time soon may still reside in the Phi Delta Kappan editor's realms of 'buncombe and ballyhoo'. That said, perhaps there is more hope of an intriguing 'breakthrough' in the integration of fast-moving developments in ability and assessment characteristic matching (CATs), in mimicking aspects of human judgement (AESs) and in sophisticated process data-related machine training for formative assessment. The power of such systems to provide appropriate and purposeful formative assessment support for learners in MOOCs and other ITSs, through personal mobile devices for example, is perhaps a little nearer with every advance in the physical technology and the underlying AI systems.

ACKNOWLEDGMENTS

The Centre for Assessment Research, Policy and Practice in Education (CARPE) is supported by a grant from Prometric Inc., a testing services provider headquartered in Baltimore, Maryland. The views expressed in the paper are solely the responsibility of the authors and have not been influenced in any way by Prometric Inc. Open access funding provided by IReL.

CONFLICT OF INTEREST

The author declares that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/jcal.12577.

DATA AVAILABILITY STATEMENT

N/A

ORCID

John Gardner https://orcid.org/0000-0002-3844-7305
Michael O'Leary https://orcid.org/0000-0002-6771-904X
Li Yuan https://orcid.org/0000-0002-7144-9441

REFERENCES

ActNext (2020) Educational Companion. https://actnext.org/research-and-projects/holistic-learning-mobile-app/

Aljohani, N. R. & Davis, H. C. (2013). Learning analytics and formative assessment to provide immediate detailed feedback using a student centred mobile dashboard. In *Proceedings of the Seventh International Conference on Next Generation Mobile Apps, Services and Technologies*, IEEE. https://www.semanticscholar.org/paper/Learning-Analytics-and-Formative-Assessment-to-a-Aljohani-Davis/4e27f92476194013d098d37b1a8e106b171726f8

Aybek, E. C., & Demirtasli, R. N. (2017). Computerized adaptive test (CAT) applications and item response theory models for polytomous items. *International Journal of Research in Education and Science*, *3*(2), 475–487.

Beigman Klebanov, B., Leong, C. W. & Flor, M. (2015). Supervised word-level metaphor detection experiments with concreteness and reweighting of examples. *Proceedings of the Third Workshop on Metaphor in NLP*, 11–20. Association for Computational Linguistics.

Beigman Klebanov, B., Madnani, N., Burstein, J. & Somasundara, S. (2014). Content importance models for scoring writing from sources. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 247-252.

Ben-Simon, A., & Bennett, R. E. (2007). Towards more substantively meaningful automated essay scoring. *Journal of Teaching, Learning and Assessment*, *6*(1), 4–44. http://www.jtla.org

Botarleanu, R. M., Dascalu, M., Sirbu, M. D., Crossley, S. A., & Trausan-Matu, S. (2018). ReadME – Generating personalized feedback for essay writing using the ReaderBench framework. In *3rd Int. Conf. on Smart Learning Ecosystems and Regional Development (SLERD 2018)* (pp. 133–145). Aalborg, Denmark.

Castro, D., McLaughlin, M. & Chivot, E. (2019) Who is winning the AI race: China, the EU or the United States? Center for Data Innovation. https://www.datainnovation.org/2019/08/who-is-winning-the-ai-race-china-the-eu-or-the-united-states/

Cicchinelli, A., Veas, E., Pardo, A., Pammer-Schindler, V., Fessl, A., Barreiros, C., & Lindstädt, S. (2018). *Finding traces of self-regulated learning in activity streams*. ACM Press. https://doi.org/10.1145/3170358.3170381

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

Cope, B., & Kalantzis, M. (2016). Big data comes to school: Implications for learning, assessment and research. *AERA Open*, *2*(2), 1–19.

Crossley, S. A., & Kyle, K. (2018). Assessing writing with the tool for the automatic analysis of lexical sophistication (TAALES). *Assessing Writing*, *38*, 46–50.

CWPA, NCTE & NWP. (2011). National Framework for success in post-secondary writing. Council of Writing Program Administrators, the National Council of Teachers of English, and the National Writing Project. http://wpacouncil.org/files/framework-for-success-postsecondary-writing.pdf

Davey, T. (2011). *A guide to computer adaptive testing systems*. Council of Chief State School Officers. https://files.eric.ed.gov/fulltext/ED543317.pdf

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

DiCerbo, K. E., & Behrens, J. T. (2014). *Impacts of the Digital Ocean on education*. Pearson.

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

ETS. (2018). GRE volumes by country. https://www.ets.org/gre/pdf/gre_volumes_by_country.pdf

Farra, N., Somsundaran, S., & Burstein, J. (2015). Scoring persuasive essays using opinions and their targets. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications* (pp. 64–74). Association for Computational Linguistics. https://www.aclweb.org/anthology/W15-0608/

García-Martínez, C., Cerezo, R., Bermudez, M., & Romero, C. (2019). Improving essay peer grading accuracy in massive open online courses using personalized weights from student's engagement and performance. *Journal of Computer Assisted Learning.*, *35*, 110–120. https://doi.org/10.1111/jcal.12316

Gartner Glossary. (2019). Big Data. https://www.gartner.com/en/information-technology/glossary/big-data

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & CAI, Z. (2004). CohMetrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments and Computers*, *36*(2), 193–202.

Gutierrez, I. & Crespo Garcia, R. M. (2012). Towards efficient provision of feedback supported by learning analytics. In *12th IEEE conference on advanced learning technologies (ICALT)*, 599–603. DOI: https://doi.org/10.1109/ICALT.2012.171

Hall, J. D., O'Connell, A. B., & Cook, J. G. (2017). Predictors of student productivity in biomedical graduate school applications. *PLoS One*, *12*(1), e0169121. https://doi.org/10.1371/journal.pone.0169121

Hauser, C. & Kingsbury, G. (2004). Differential item functioning and differential test functioning in the Idaho Standards Achievement Tests for Spring 2003. https://files.eric.ed.gov/fulltext/ED491248.pdf

Heilman, M., Cahill, A., Madnani, N., Lopez, M., Mulholland, M. & Tetreault, J. (2014). Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 174-180. https://www.aclweb.org/anthology/P14-2029.pdf

Hewlett Packard. (2012) The Hewlett Packard Foundation: Automated Essay Scoring Competition (to develop an automated scoring algorithm for student-written essay). https://www.kaggle.com/c/asap-aes

Jansen, R. S., van Leeuwen, A., Jansen, J., Conijn, R., & Kester, L. (2019). Supporting learners' self-regulated learning in massive open online courses. *Computers and Education*, *146*, 103771. https://doi.org/10.1016/j.compedu.2019.103771

Jarvela, S., Gasevic, D., Seppanen, T., Pechinizky, M., & Kirschner, P. (2020). Bridging learning sciences, machine learning and affective computing for understanding cognition and affect in collaborative learning. *British Journal of Educational Technology*, *51*, 2391–2406. https://doi.org/10.1111/bjet.12.917

Ju, G.-F. N., & Bork, A. (2005). The implementation of an adaptive test on the computer. In *Proceedings of the fifth IEEE international conference on advanced learning technologies* (pp. 822–823). IEEE Computer Society. https://doi.org/10.1109/ICALT.2005.152

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046.

LAK (2011). 1st International Conference on Learning Analytics and Knowledge. https://tekri.athabascau.ca/analytics/

Lerche, T., & Kiel, E. (2018). Predicting student achievement in learning management systems by log data analysis. *Computers in Human Behavior*, *89*, 367–372. https://doi.org/10.1016/j.chb.2018.06.015

Long, P., & Siemens, G. (2011). Penetrating the fog. *Educause Review*, *46*(5), 31–40.

Lord, F. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer assisted instruction, testing and guidance* (pp. 139–183). Harpur and Row.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Martin, F., & Ndoye, A. (2016). Using learning analytics to assess student learning in online courses. *Journal of University Teaching & Learning Practice*, *13*(3), 1–20.

McKinsey and Co. (2011). *Big data: The next frontier for innovation*. Competition and Productivity. https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%

20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_exec_summary.ashx

Miller, C., & Stassun, K. (2014). A test that fails. *Nature*, *510*, 303–304.

Molenaar, I., Horvers, A., & Baker, R. S. (2019). What can moment-by-moment learning curves tell about students' self-regulated learning? *Learning and Instruction*, *72*, 101206. https://doi.org/10.1016/j.learninstruc.2019.05.003

Moneta-Koehler, L., Brown, A. M., Petrie, K. A., Evans, B. J., & Chalkley, R. (2017). The limitations of the GRE in predicting success in biomedical graduate school. *PLoS One*, *12*(1), e0166742. https://doi.org/10.1371/journal.pone.0166742

NCTE. (2013). *Position statement on machine scoring*. National Council of Teachers of English. http://www2.ncte.org/statement/machine_scoring/

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

Partnership on AI. (n.d.). (80 major corporations, inc. BBC, Apple, Google and Amazon, concerned with the best practice use of AI) Report on Algorithmic Risk Assessment Tools in the US Criminal Justice System. https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/

Pearson. (n.d.). *Intelligent Essay Assessor (IEA) fact sheet*. https://images.pearsonassessments.com/images/assets/kt/download/IEA-FactSheet-20100401.pdf

Perelman, L. (2012a). Mass marketing assessment of writing is bullshit. In N. Elliott & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. white* (pp. 425–437). Hampton Press.

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

Perelman, L. (2014). When 'state of the art' is counting words. *Assessing Writing*, *21*, 104–111.

Perelman, L. (2018). *Interview on his babel generator Tovia Smith podcast: More states opting to 'robo-grade' student essays by computer*. National Public Radio. https://www.npr.org/2018/06/30/624373367/more-states-opting-to-robo-grade-student-essays-by-computer?t=1564600373190

Reckase, M. D., Ju, U., & Kim, S. (2019). How adaptive is an adaptive test: Are all adaptive tests adaptive? *Journal of Computerized Adaptive Testing*, *7*(1), 1–14.

Rudner, L. M. (1998). An on-line, interactive Computer Adaptive Testing mini-tutorial. http://echo.edres.org:8080/scripts/cat/catdemo.htm

Somasundaran, S., Lee, C. M., Chodorow, M., & Wang, X. (2015). Automated scoring of picture-based story narration. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications* (pp. 42–48). Association for Computational Linguistics. https://doi.org/10.3115/v1/W15-06

Spector, J. M., Ifenthaler, D., Sampson, D., Yang, L., Mukama, E. & 12 additional authors (2016) Technology enhanced formative assessment for 21st century learning. *Educational Technology and Society*, *19*(3) 58–71.

Stocking, M. L., Smith, R., & Swanson, L. (2000). *An Investigation of approaches to Computerising the GRE subject tests*. Education and Testing Services. https://doi.org/10.1002/j.2333-8504.2000.tb01827.x

Tempelaar, D. T., Cuypers, H., van de Vrie, E., Heck, A. & van der Kooij, E. (2013) Formative assessment and learning analytics. In *LAK '13: Proceedings of the Third International Conference on Learning Analytics and Knowledge*, Leuven Belgium, 205–209. DOI:https://doi.org/10.1145/2460296.2460337

Thille, C., Kizilcec, R., Piech, C., Halawa, S. A., & Greene, D. K. (2014). The future of data-enriched assessment. *Research and Practice in Assessment*, *9*, 5–16.

Tuomi, I., Cabrera Giraldez, M., Vuorikari, R., & Punie, Y. (2018). *The impact of artificial intelligence on learning, teaching, and education*. Publications Office of the European Union. https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/impact-artificial-intelligence-learning-teaching-and-education

UNESCO. (2019). Artificial intelligence in education: Challenges and opportunities for sustainable development. Education Sector, United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000366994/PDF/366994eng.pdf.multi

von Davier, M. (2019). Training Optimus prime, M.D.: Generating medical certification items by fine-tuning OpenAI's gpt2 transformer model. https://arxiv.org/abs/1908.08594

Walker, E., Wong, A., Fialko, S., Restrepo, M. A., & Glenberg, A. M. (2017). EMBRACE: Applying cognitive tutor principles to Reading comprehension. In E. André, R. Baker, X. Hu, M. Rodrigo, & B. du Boulay (Eds.), *Artificial intelligence in education. 18th international conference, AIED 2017 Wuhan, China, June 28 – July 1, 2017 proceedings* (pp. 578–581). Cham.

Wise, S. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing*, *2*(1), 1–17.

How to cite this article: Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?'. *Journal of Computer Assisted Learning*, *37*(5), 1207–1216. https://doi.org/10.1111/jcal.12577