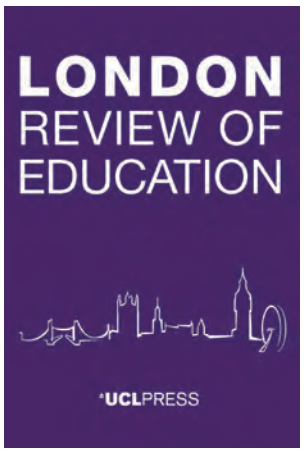

UCLPRESS

LONDON REVIEW OF EDUCATION

e-ISSN: 1474-8479

Journal homepage:
https://www.uclpress.co.uk/pages/london-review-of-education

Rise of the machines? The evolving role of AI technologies in high-stakes assessment

Mary Richardson and Rose Clesham

How to cite this article
Richardson, M. and Clesham, R. (2021) 'Rise of the machines? The evolving role of AI technologies in high-stakes assessment'. *London Review of Education*, 19 (1), 9, 1–13. https://doi.org/10.14324/LRE.19.1.09

Submission date: 4 October 2019
Acceptance date: 29 September 2020
Publication date: 10 March 2021

Peer review
This article has been peer-reviewed through the journal's standard double-blind peer review, where both the reviewers and authors are anonymized during review.

LONDON REVIEW OF EDUCATION

Richardson, M. and Clesham, R. (2021) 'Rise of the machines? The evolving role of AI technologies in high-stakes assessment'. *London Review of Education*, 19 (1), 9, 1–13. https://doi.org/10.14324/LRE.19.1.09

Rise of the machines? The evolving role of AI technologies in high-stakes assessment

Mary Richardson* – *UCL Institute of Education, UK*
Rose Clesham – *Pearson Education Ltd, UK*

Abstract

Our world has been transformed by technologies incorporating artificial intelligence (AI) within mass communication, employment, entertainment and many other aspects of our daily lives. However, within the domain of education, it seems that our ways of working and, particularly, assessing have hardly changed at all. We continue to prize examinations and summative testing as the most reliable way to assess educational achievements, and we continue to rely on paper-based test delivery as our modus operandi. Inertia, tradition and aversion to perceived risk have resulted in a lack of innovation (James, 2006), particularly so in the area of high-stakes assessment. The summer of 2020 brought this deficit into very sharp focus with the A-level debacle in England, where grades were awarded, challenged, rescinded and reset. These events are potentially catastrophic in terms of how we trust national examinations, and the problems arise from using just one way to define academic success and one way to operationalize that approach to assessment. While sophisticated digital learning platforms, multimedia technologies and wireless communication are transforming what, when and how learning can take place, transformation in national and international assessment thinking and practice trails behind. In this article, we present some of the current research and advances in AI and how these can be applied to the context of high-stakes assessment. Our discussion focuses not on the question of whether we should be using technologies, but on how we can use them effectively to better support practice. An example from one testing agency in England using a globally popular test of English that assesses oral, aural, reading and written skills is described to explain and propose just how well new technologies can augment assessment theory and practice.

Keywords: artificial intelligence, assessment, high-stakes testing, language tests

Introduction

This article considers how research and advances in artificial intelligence-led technologies are viewed and understood in the context of high-stakes assessment. We discuss these issues using an example from one testing agency in England which has established a globally popular test of English that assesses oral, aural, reading and written skills. The test uses artificial intelligence (AI) technologies to support practical elements of test delivery, and it also employs AI to determine grading and awarding outcomes. Despite there being a prevalence of AI and related technologies in everyday life, within the context of educational research and within education settings, such as schools and colleges, the use of technology (particularly technology including AI) has been slow to evolve. Within the domain of educational assessment, it is still considered

*Corresponding author – email: Mary.richardson@ucl.ac.uk

the new kid on the block. In many senses, this is surprising because there is obvious value in harnessing contemporary computing power to facilitate AI and automated machine decision making in the processing of data such as exam results. This was never so relevant as in 2020, when the entirety of the national high-stakes testing systems in England ground to a halt due to the COVID-19 pandemic. The simple fact that students could no longer sit in an exam hall in front of paper-based examinations meant that the entire system had to be revised. It is the reliance on antiquated systems of testing that led to the slow unravelling of the now infamous examinations debacle (see, for example, Barton, 2020; McDonagh, 2020).

We might expect that the functionality of AI technologies alone would be a verifiable means to improve the accuracy and reliability of assessment practice such as high-stakes national and international tests. However, it seems that there are several issues at play here. First, there is concern about the efficacy of reliance on automated systems to facilitate authentic assessments of student learning. For example, would even the best AI technology we have be able to evaluate a student's answer in a nuanced manner taking into account the broad range of insight that an experienced examiner will draw upon? Second, there is a more pragmatic issue: the fear of machines taking over a human role (Zawacki-Richter et al., 2019) – could the mechanization of this aspect of education practice potentially leave experienced assessors redundant?

To address these issues, our discussion relates to the nature of trust in the use of AI technology in assessment practice and considers one question: How do we make it work effectively for our purposes?

How do we make AI technology work effectively for our purposes?

There are several international testing agencies who assess millions of English as a second language (ESL) students each year on receptive (listening and reading) and productive (speaking and writing) skills for the extremely competitive admissions to English-speaking universities. Such goals are responsible for making tests high stakes (Chapelle and Chung, 2010), and prospective candidates rely on them to access life-changing opportunities such as citizenship or educational opportunities. Where a test holds such intrinsic power, we need to be sure that the skills are being tested with high validity and reliability and that we are able to obtain accurate scoring and fast, informed feedback. Confidence is at the heart of validity in high-stakes testing, and while it is not always considered, it is, in fact, critical to the value of assessments that wield such influence.

Concerns over the use of automated marking technologies (Falkner et al., 2014) often cite fear of a loss of the human touch as potentially harmful to the assessment process, but we propose that in the majority of cases, the opposite may be true. The use of AI technologies allows the judgement of hundreds of human assessors to work in unison and within classrooms. Indeed, used well, AI can potentially enable teachers to spend *less* time marking and *more* time focusing on teaching and learning. Recent events, such as the inability in England to present national examinations in a form other than on paper, have begun to unmask well-hidden inadequacies that underpin national testing systems – we discuss this further in the next section.

The use of AI per se has historically been viewed as something that should invoke caution. In 1920, the Czech playwright Karel Čapek released his script for a futurist view of the world; in *R.U.R. (Rossum's Universal Robots)* (Koreis, no date), the

audience was presented with a vision of the year 2000, a world dominated by robots, but, as the story unfolds, we see the machines evolve a capacity for emotion, and thus some 'hope' for the future. The uneasy relationship between the human condition and robotic/artificial technology continues to be a popular theme presented alongside serious discussions on the ways in which new technologies influence day-to-day living. However, since the 1950s, our curiosity about the capacity of machine learning and AI technologies has also evolved into a respected and multidimensional domain within scientific research and development (Bibel, 2014; Bostrom and Yudkowsky, 2014). In 2020, we rarely think about the many AI-led technologies that guide the day-to-day aspects of life, such as targeted advertising on our mobile phones or satellite navigation systems or managing the weekly shop. It is perhaps the rather mundane nature of these kinds of AI intrusions that is precisely why they are no longer visible to us; whereas when they have an effect on something more important, such as educational outcomes, they grab our attention.

Our perception of new technologies is shaped by how they impact our lives, and education domains are particularly idiosyncratic because we tend to trust those we *know* as educational enactors (Corrigan and Chapman, 2008). For example, we believe that our teachers are more trustworthy than a machine when it comes to an important practice such as assessment. However, the example of A levels in England reveals our 'flexible' attitude to trust in teacher judgement, because usually our students rely on national test outcomes that are standardized using statistical algorithms to model the data provided from markers via examination boards. In 2020, the government's wish to reject teacher-estimated grades alone and, initially at least, to insist that students must accept standardized grades generated by a statistical algorithm led to national demonstrations and the eventual retreat of government to allow students a choice of their 'best' outcome (Richardson, 2020). What was not often discussed was the fact that every year, the data that provide the evidence for awarding grades for A levels and GCSEs is always modelled using statistics in order to consider what the comparable standard might be. These data are generally considered alongside discussions with expert examiners (the human input) to determine the outcomes at grade boundaries (see Assessment and Qualifications Alliance, 2020). There are technological interventions that already form part of the awarding process and practice – for example, online marking and modelling of awarding data – but we only tend to consider their efficacy when there is a crisis. In terms of assessment, one of the most pressing issues highlighted in summer 2020 was the fact that assessment is not a simple process, nor does it provide us with precise outcomes. This knowledge is unsettling in a culture where an examination result based on paper-based test experiences is reified and considered the gold standard; expecting people to alter how they think about this public endeavour is a challenge indeed.

Assessment technologies

Automated scoring, including the use of AI, is now integral to the latest education technology innovations, particularly in the field of both formative and summative assessment practice. AI developers make a range of claims for its integrity and its application (see, for example, Moon and Pae, 2011; Pinot de Moira, 2013; Bridgeman, 2013), and they generally agree that:

- it speeds up marking times
- it removes/reduces human bias
- it is as accurate and at least as reliable as human markers.

Based on these assertions, it appears that some AI-led technology has the potential to challenge the status quo in existing educational assessment practices, particularly in how we mark and provide feedback on student work. This claim may be surprising to some, given that the perception of formative assessment modes is characterized by personal responses that are reliant on text or verbal feedback (Sadler, 1989). However, AI has a proven record in formative assessment use, particularly in the US (see, for example, Attali, 2013; Bridgeman, 2013; Luckin, 2017), and, as Whithaus (cited in Shermis and Burstein, 2013: vii) claims, 'Pretending that software systems, and particularly software agents that provide feedback, are not parts of students' writing processes, not parts of a broader ecology of available writing tools in the second decade of the 21st century, is simply naïve.' Whithaus (2006) adds that it is time to talk more about the use of technologies and to open up more public discussions about their use and value in education. We agree, particularly given that they are so widely used in high-stakes assessment in English language testing, as we explain later in this article.

However, both within and beyond the education sector, there is some hesitancy in accepting new, AI-led practice, particularly in countries or jurisdictions with long-established paper-based testing histories. Paradoxically, within technology-enabled assessment systems, few people actually ask the question 'Does it actually *work*?'; rather, the more common question posed is 'OK, *how* does it work for our purposes?' This is where language testing is leading the way in the use of particular technologies to support not only assessment, but also student learning, and those engaged in their development are keen to promote the value of moving above and beyond the established ways of testing which rely on paper.

In a world already transformed by technology in the way people communicate, work and live their daily lives, most educational assessment in the UK has hardly changed at all (Timmis et al., 2016). Despite continuous discourses that promote AI technologies, we continue to not trust them in high-stakes assessments.

Across all areas of education, within the sciences, society and politics, and throughout commerce, it is not simply that new technologies are boosting our abilities; they are also actively *influencing and guiding* them (Natarajan et al., 2017; Paschen et al., 2020), for better or, indeed, for worse. Using neural networks and deep learning, AI is increasingly being employed in higher education selection and screening procedures to ensure that students and workers have a construct representative, fair, valid and reliable assessment for migrant entry or for study purposes across the world. Bearing this in mind, perhaps we should be proactive in understanding just how they influence what we do, because they are a tool and, as Shermis and Burstein (2013) and Shaw (2008) argue, part of the educational journey with AI technologies is to consider how to create sustainable automated systems that support and serve us.

Bridle (2018: n.p.) claims that, in many ways, technologies are 'extensions of ourselves, codified in machines and infrastructures, in frameworks of knowledge and action', but he adds that they do not have all of the answers we seek. It is important to understand the limitations as well as the potential of current technology. Such issues have been brought into sharp focus in 2020 as the impact of the COVID-19 pandemic has challenged national testing and examination systems around the world simply because students could not sit their exams in the usual way. The validity of results for the very highest of high-stakes exams is under the spotlight and, in England, so is the reliance on an algorithm (not using any AI) used to calculate outcomes based on testing centre predictions and then standardized using comparable outcomes (Pearson, 2017). This approach was problematic in the extreme. The final grades triggered a national outcry from students, teachers and parents demanding fairer

outcomes, given that 39 per cent found their calculated grades had then been reduced so that the comparable outcome percentages could be maintained compared to 2019. Following several U-turns from the UK government, the grades awarded were not those generated by the algorithm and adjusted, but those predicted by teachers. These events produced a great deal of fodder for the news and social media (for a useful summary, see Richardson, 2020), but, more importantly, the debates highlighted a broadly weak understanding of just what it is that algorithms do. Even the prime minister demonstrated his ignorance by calling it a 'mutant algorithm' (Stewart, 2020), as if it had some mind of its own or was reliant on some AI technology, whereas this model was something much more straightforward. The outputs from this kind of model are only as good as the data that go in, and with a lack of the most crucial indicators – the exam results – the expectation that outcomes would be accurate was wildly optimistic. The fallout from the debacle finally resulted in a parliamentary inquiry (UK Parliament, 2020a, 2020b), and there will be more to follow as we look ahead to planning for national assessments in 2021.

Reliance on paper-based examinations for national high-stakes tests has proven to be fallible in the face of a global pandemic, as test takers were unable to sit traditional examinations due to enforced isolation. There has been an increase in the interest and use of self-proctored online examinations, where test takers sit tests in their own homes and log into secure environments. Of course, as with all high-stakes tests, the continued issue is trust in the technology, but, as we will discuss in the next section, AI potentially affords a range of ways to engender trustworthiness.

'Good technology' and assessment in the UK

The notion of good technology and assessment can be conceptualized in three separate yet intertwined issues in the constructs, delivery mechanism and efficiency/reliability of national assessment systems. When Ken Boston (2005), Chief Executive of the Qualifications and Curriculum Authority (QCA) – the executive non-departmental public body of the Department for Education in the United Kingdom – announced that he expected on-screen assessments for all new qualifications by 2009, and that e-assessment should be a routine provision in this country by that time, he and the QCA had a number of motivations. It was clear then, and has continued to be the case ever since, that there is a need to incorporate twenty-first-century technology into education generally, and into assessment specifically. Ripley (2004) was already arguing for modernization of the logistics of examination management in this country. As the number of externally marked tests and exams increased exponentially, so the cottage industry of scripts being sent through the post to the doorsteps of examiners was deemed to be, at the least, insecure and, at most, costly in terms of both time and people. These systems are beginning to change as more marking is undertaken on screen.

We have by no means reached a point where we are making appropriate use of good technology, but the motivation for change should not be seen as the sole driving force for assessment changes. One should never underestimate the importance of quality, robustness and efficiency of assessment operational systems in the perceived success or failure of national assessments – the public face of assessment is critical in terms of engendering trust and confidence. The numerous debacles of test delivery failures (see, for example, Isaacs, 2014) and the continued reporting of missing or stolen exam scripts (Curtis, 2005; Westfield, 2016) also exemplify the continued challenges to high-stakes national tests. The more public

discussions and promotion of such issues in national testing led to claims of a need to explore the way in which new technologies might provide a more secure structure for delivery and processing.

In 2007, Ken Boston continued to argue for serious consideration of how new technologies might facilitate improved assessment systems, offering, for example, personalized learning, assessment on demand and rapid feedback of results and data to assist exploration of assessment reliability and validity related to on-screen assessments. Although it might be assumed that curriculum and constructs drive assessment change rather than those of efficiency, it might be argued that the latter reason initialized the movement towards computer-based assessment. The aspirations and expectations expressed in 2005 and 2007 were ambitious in the extreme and, unsurprisingly, have not been met. This is in part due to the lack of IT infrastructure in schools across England (Brown et al., 2013), but also largely due to comparability and equivalence issues, as dual modalities of testing (computer- and paper-based) have presented insurmountable problems for awarding bodies and regulators.

Therefore, despite the claims made almost two decades ago, the expected e-assessment revolution has remained largely unfulfilled, and where there has been evidence of e-assessment in practice, we see just 'paper-behind-glass', that is, paper-based assessment captured and represented on a screen. While addressing some of the equivalence and efficiency issues, simply presenting a paper-based test on computers offers no affordances in terms of the use of technology to provide stimulating and authentic assessment opportunities. An early example of problematic electronic test items is discussed by Richardson et al. (2002), who conducted observations of students in schools in England taking a range of new problem-solving tests on computers. They found that the interactive nature of items distracted students and that their attention was easily diverted to construct irrelevant aspects of the test designs featuring cartoon images. What is important to note is that such design features provided little more than a representation of the paper-based test format, rather than a new test-taking experience that was tailored to the delivery technology rather than the expected learning.

The most basic computer-based test assessments might include a mixture of multiple-choice options and open responses sent to human markers. Others include what are described as *constrained test forms* (see Liu et al., 2005) – those consisting of all fixed-form question types such as multiple-choice or drag-and-drop answering functions. While e-assessments have provided logistical support in providing more flexible and technology-assisted testing opportunities, they have not addressed more assessment validity related questions, particularly associated with providing construct representative, authentic, reliable, fast and fair assessments on a global basis.

However, amid the general stasis of the early forms of e-assessment in UK education, there have been pockets of innovative technologies involving the use of AI. These have contributed to some transformations, not only in the manner in which learning takes place, with personalized, instantaneous and engaging formative assessment experiences (Waring and Evans, 2015), but also in how high-stakes summative assessments can be delivered globally. The global context is important, and this is reflected in the use of more electronic testing technologies in the international large-scale assessments (ILSAs), such as PISA, TIMSS and PIRLS (OECD, 2020; International Association for the Evaluation of Educational Achievement: www.iea.nl/). Originally set up to share information about processes and practices in educational policy around the world (Torney-Purta and Amadeo, 2013), the power of these sets of big data has been augmented by new modes of analysis, electronic modes of

testing, translation and the use of cloud technologies to hold and share information. However, similarly to other areas of mass testing, the ILSAs have been slow to adopt new technologies for the actual delivery and presentation of tests themselves. In 2019, the TIMSS cycle trialled the use of tablet-based tests in some participant countries (Cockle and Sibberns, 2019), and the results of these data compared to data for students who took the tests on paper formed part of the international reporting in December 2020 (Richardson et al., 2020). On a positive note, technology has facilitated instant communication, and greater capacity to share information and ideas, and has arguably brought us closer to one another (particularly during the lockdown of 2020, when COVID-19 limited movement in public spaces). While such experiences constrain us in some ways, in terms of educational settings, they have forced open new doors to spaces for sharing ideas about education policy and practice.

Case study: AI in tests of English language

Within the field of educational assessment, the curriculum area in which AI has emerged is language testing, and this has not been a random choice for test developers. Every year, millions of candidates (British Council, 2020; Pearson Vue, 2021) are assessed on both receptive (listening and reading) and productive (speaking and writing) skills for entry into English-speaking universities or professions, or for citizenship. Given the opportunities available to those who succeed, these tests are very high stakes, and prospective candidates are competing in a global context. Therefore, developing a test that is high in validity and reliability is central to the success of these assessments, and given the global nature of study and work, there is a need for tests to be reliable, fast and accessible.

In language testing, response types for receptive skills often take the form of multiple choice or constrained response (Liu et al., 2005), whether the mode of assessment is paper-based or on screen. However, the assessment of speaking and writing skills, where all responses will be unique, cannot be assessed using these fixed form types, and so the default marking methodology requires the use of thousands of human markers, attempting to maintain a global standard and achieve acceptable levels of marking reliability. As might be expected, the reliability of assessing speaking and writing skills (that is, open-ended productive skills) has always been a significant challenge on a local and national level (Brooks, 2012; Rhead and Black, 2018). To manually collect these types of responses and get them marked, and scaled, reliably is logistically difficult, time consuming and prone to standardization issues. Of course, assessment history has demonstrated that the solution to issues surrounding logistical problems and low reliabilities of productive skills have often resulted in their removal from the assessed construct. A good example of this is the removal of speaking skills assessment from the national curriculum in England, mainly due to the low reliability of the assessment outcomes (Stobart, 2009). No one had argued that speaking was not an essential part of the English language construct taught in schools; however, this essential skill was jettisoned from formal assessment. What influences such decisions are, argues Wiliam (2001), the consequences of narrowing the curriculum for assessment construct purposes. Essentially, the very slender assessed construct becomes the focus for what constitutes the taught construct – the assessment *is* the curriculum. Once this happens, the assessed curriculum becomes more important than the intended curriculum (Stobart, 2008) and, in the end, instead of making the important measurable, we make the measurable important. The logistical and reliability related reasons for removing sections of a construct are a case in point for the potential that

AI can provide in enabling complete construct representation in language testing, with incredibly high levels of both internal and marking reliabilities.

So, we return to our question 'How does AI work for our assessment purposes in English language testing?' It is important that stakeholders understand the principles of the application of AI in educational assessment. While there are technical complexities involved, it is incumbent on test providers to open up the black box in order to present a compelling validity argument for its use in speaking and writing assessments. The next section will outline this.

Automated scoring

Automated scoring of spoken language requires three models to be developed: an acoustic model, a language model and a scoring model. Acoustic models are speech recognizers, identifying each phoneme or sound (Young, 2001), and while there was a time when acoustic models were seen as 'futuristic', they are very much a part of daily life. All smartphone users have access to acoustic models as a matter of course and use them for a wide range of practical tasks, such as passwords or transferring oral speech into text. This means that it is vital that a good acoustic model is trained using a wide range of accents and pronunciation types in order to be able to recognize the broadest representation for its use.

The language model is then developed by training the AI system on every spoken task item. This requires trialling all items on a broad representative sample globally, and at least 400 trial responses are used to train the AI system on each item. The next step is to incorporate the scoring model by transcribing all the spoken responses and the marks they were given by a team of expert human raters (all items are double marked as a minimum). Only once these steps are complete can we say that the AI system is 'trained'. At this point, it is important to note that the human touch is still required because these data are validated by scoring at least four hundred new items of each prompt that have also been transcribed and marked by expert raters. The reliability coefficients between the scores obtained by the AI system and humans is in the region of 0.96 agreement (Pearson, 2019: 6). By any terms, this is incredibly high (see Viera and Garrett, 2005), and any item that does not achieve this reliability score is removed from the test. Therefore, although AI is a machine prediction of a score, the benchmark for the inclusion of any particular item prompt is *human* judgement.

The automated marking of open-ended writing works in a similar way. Pearson's scoring engine, Intelligent Essay Assessor (IEA), evaluates meaning using a natural language processing technique called latent semantic analysis (LSA) (Foltz et al., 2013). With LSA, each word, sentence and passage becomes what is called a *vector* in relation to a multidimensional semantic space. For example:

> Surgery is often performed by a team of doctors.
> On many occasions, several physicians are involved in an operation.
> (Foltz et al., 2013: 78)

In this example, the two sentences contain no words in common, but their meanings are approximately the same based on the contexts of the words used. The words 'physicians' and 'doctors' appear in similar contexts in English, so, in an LSA vector space, these two sentences would describe effectively the same vector because their underlying meaning is the same. Behind this sits the content-based scoring, a vital 'background' model using an enormous corpus to evaluate a newly submitted essay

or summary assignment. When scoring prompts specific traits, IEA compares the incoming essay with all known scored essays from the training set and determines the new essay's vector proximity (called a cosine) to other known pre-scored essay vectors in the semantic space. LSA variables are used to predict not only content, word choice and task completion, but also organizational traits such as sentence fluency and essay coherence. Again, the reliability of the machine writing scores compared to humans is strong – in the region of 0.88 agreement (Pearson, 2019: 4).

The use of speech and writing automated marking technologies allows speaking and writing to be assessed for formative and/or summative purposes across a broad range of identified traits, and this provides opportunities for students to obtain detailed feedback on their strengths and weaknesses, both inside and outside of a formal classroom. Teachers can use these technologies as part of a blended approach to teaching and learning; they can use the automated feedback to augment personal formative assessment opportunities with their students.

Although AI assessment can be accused of being rather opaque, in some respects, machine scores are more transparent than human judgements. Human raters evaluate language samples, refer to scale descriptors and apply judgement and experience to assign a final score, but there is often no quantifiable way of measuring how they weight and combine the various pieces of information in an essay. In reality, we rely on a lot of experiential judgements and knowledge on the part of examiners (Newstead and Dennis, 1994; Greatorex and Bell, 2008; Johnson et al., 2012). It is hard to explain that you know what an assessment is worth when judged against some criteria, but there is definitely an element of 'gut feeling' in marking. In contrast, it is possible to achieve something replicable with machine scoring, with every piece of data analysed, and its precise weighting in the scoring model is verifiable in the machine algorithms. As Bernstein et al. (2010) claim, scoring models are data-driven and verifiable, often in ways that human scoring is not. AI systems are quicker, less prone to error and capable of assessing a number of key traits simultaneously; importantly, they do not carry any inherent halo or negative effects (Dennis, 2007; Beltrama and Hanink, 2019), and they treat all responses fairly – as long as the AI systems are trained on sound, representative samples.

This is not to say, however, that AI can be used in all assessment areas. The argument here is not binary: that there should be all or no AI. Where appropriate, AI is as good as, or indeed better than, human assessors, and it can act as an invaluable aid to logistics and constructs of the assessment industry as a whole, and to teachers in the classroom. Reflecting on the technical challenges that beset the national testing systems in England in 2020, it appears that a tentative move to considering how some automation of processes and application of AI technologies could be used in future is already happening. But we should also be aware that there is no easy way to manage assessment and, of course, there may be areas of the curriculum where AI will never be the right assessment tool. Whether AI will ever be able to predict creative thinking, or, indeed, as envisaged by Čapek's robots of 1920, emotional responsiveness, remains to be seen.

Conclusion

The reliance on paper-based testing holds fast in England (and many other countries, for that matter), and due to the fact that high-stakes, secure assessment (such as tests of language competency) has become global currency, the consequences often disappoint stakeholders in terms of inauthentic, invalid, non-construct representative

qualifications or awards. We could not wish for a better example than 2020 in terms of an inflexible assessment system that is unable to respond to trauma of the kind caused by COVID-19. In contrast, the rapid developments in emerging technologies in hardware and software have the potential to afford students and teachers ever advancing opportunities to be assessed in authentic, broad ways that are fast, efficient, reliable and, when needed, secure. The use of virtual reality, augmented reality and video capture has the potential to allow immersion into areas of the curriculum and occupations not possible in traditional learning and assessment methods.

The use of AI in language assessment has proved to be a game changer internationally and, as we have discussed above, AI technologies are being used to develop specialist tests of English language and to outline their effectiveness in terms of validity and reliability of both test delivery and scoring. We recognize that there is no perfect assessment, and we are well aware of the fact that all types of test are subject to error, but the example described here provides evidence of how AI technology can assure the test taker of a reliable result.

In many ways, instead of debating whether AI *should* be incorporated more into formative and summative assessments, we should perhaps be asking why this *would not* be the case? In a recent survey asking students about their preferred mode of international language test, four of the top five reasons were AI related (Eduworld, 2018). The results were focused on security, speed and accessibility.

The use of AI technologies in assessment has to be context related to ensure the validity of its use. Simply putting former paper-based items into an AI-led environment does not guarantee an enhanced test-taking experience, nor does it add anything of pedagogical value to the core constructs of the assessment. Returning to Bridle's (2018: n.p.) cautionary discourse on AI reminds us that:

> Computational systems, as tools, emphasise one of the most powerful aspects of humanity: our ability to act effectively in the world and shape it to our desires. But uncovering and articulating those desires, and ensuring that they do not degrade, overrule, efface, or erase the desires of others, remains our prerogative.

The use of AI in language assessment has proved to be a game changer internationally. However, all stakeholders in the test-taking process need to be aware of the need for improved literacy in this new test-taking landscape. Assessment literacy is not a new idea (Stiggins, 1995; Popham, 2009), but assessment literacy employing new technologies is still a work in progress, and one that requires further research because the future of assessment in general may indeed be transformed through the technology of AI.

Notes on the contributors

Mary Richardson is Associate Professor of Education at UCL Institute of Education, UK. She leads the MA in Education (Assessment) and supervises doctoral candidates researching assessment, ethics and citizenship. Her research interests include assessment, ethics and citizenship. She is writing a book about confidence in assessment. She is Principal Investigator for the research strand of the TIMSS2019 project for England. She has recently started a three-year research project investigating the user experience of AI tests of language.

Rose Clesham is Director of Academic Standards and Measurement, working in Global Assessment at Pearson. She is an expert in assessment design and research, implementing national and international alignment and benchmarking studies, working

on OECD PISA assessments. She co-wrote the 2015 Scientific Literacy Framework. Her research interests include the emerging development of e-assessment and artificial intelligence, and ongoing educational strategies and policy. Rose is a visiting associate professor at UCL Institute of Education, UK.

Declarations and conflict of interests

The authors declare no conflict of interest with this work.

References

AQA (Assessment and Qualifications Alliance) (2020) 'How grades are awarded'. Accessed 25 September 2020. www.aqa.org.uk/about-us/what-we-do/getting-the-right-result/how-exams-work/making-the-grades-a-guide-to-awarding.

Attali, Y. (2013) 'Validity and reliability of automated essay scoring'. In M.D. Shermis and J. Burstein (eds), *Handbook of Automated Essay Evaluation*. New York: Routledge. https://doi.org/10.4324/9780203122761.ch11.

Barton, G. (2020) 'Coronavirus: Will students pay for a lack of a Plan B?' *Tes*, 18 September. Accessed 25 September 2020. www.tes.com/news/coronavirus-schools-will-students-pay-governments-lack-plan-b.

Beltrama, A. and Hanink, E. (2019) 'Marking imprecision, conveying surprise: *Like* between hedging and mirativity'. *Journal of Linguistics*, 55 (1), 1–34. https://doi.org/10.1017/S0022226718000270.

Bernstein, J., Van Moere, A. and Cheng, J. (2010) 'Validating automated speaking tests'. *Language Testing*, 27 (3), 355–77. https://doi.org/10.1177/0265532210364404.

Bibel, W. (2014) 'Artificial intelligence in a historical perspective'. *AI Communications*, 27 (1), 87–102. https://doi.org/10.3233/AIC-130576.

Boston, K. (2005) 'Assessment, reporting and technology: System-wide assessment and reporting in the 21st century'. In *Tenth Annual Roundtable Conference: Strategy, technology and assessment*. London: Qualifications and Curriculum Authority.

Boston, K. (2007) 'Tipping points in education and skills'. Speech to QCA Annual Review, London.

Bostrom, N. and Yudkowsky, E. (2014) 'The ethics of artificial intelligence'. In K. Frankish and W. Ramsey (eds), *The Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge University Press, 316–34. https://doi.org/10.1017/cbo9781139046855.020.

Bridgeman, B. (2013) 'Human ratings and automated essay evaluation'. In M.D. Shermis and J. Burstein (eds), *Handbook of Automated Essay Evaluation: Current applications and new directions*. New York: Routledge, 221–33. https://doi.org/10.4324/9780203122761.ch13.

Bridle, J. (2018) 'Rise of the machines: Has technology evolved beyond our control?'. *The Guardian*, 15 June. Accessed 1 November 2019. www.theguardian.com/books/2018/jun/15/rise-of-the-machines-has-technology-evolved-beyond-our-control-.

British Council (2020) 'British Council and assessment'. Accessed 1 December 2020. www.britishcouncil.org/exam/aptis/research/projects/assessment-literacy/introducing-language-assessment-0.

Brooks, V. (2012) 'Marking as judgment'. *Research Papers in Education*, 27 (1), 63–80. https://doi.org/10.1080/02671520903331008.

Brown, N., Kölling, M., Crick, T., Jones, S.P., Humphreys, S. and Sentance, S. (2013) 'Bringing computer science back into schools: Lessons from the UK'. In *Proceedings of the 44th ACM Technical Symposium on Computer Science Education* (*SIGCSE 2013*). New York: Association for Computing Machinery, 269–74.

Chapelle, C.A. and Chung, Y.-R. (2010) 'The promise of NLP and speech processing technologies in language assessment'. *Language Testing*, 27 (3), 301–15. https://doi.org/10.1177/0265532210364405.

Cockle, M. and Sibberns, H. (2019) 'eAssessment system for TIMSS 2019'. In M.O. Martin, M. von Davier and I.V.S Mullis (eds), *Methods and Procedures: TIMSS 2019 technical report*. Accessed 12 January 2021. https://timssandpirls.bc.edu/timss2019/methods/chapter-4.html.

Corrigan, M.W. and Chapman, P.E. (2008) 'Trust in teachers: A motivating element to learning'. *Radical Pedagogy*, 9 (2). Accessed 12 January 2021. https://radicalpedagogy.icaap.org/content/issue9_2/Corrigan_Chapman.html.

Curtis, P. (2005) 'GCSE exam papers go missing'. *The Guardian*, 9 May. Accessed 3 September 2020. www.theguardian.com/education/2005/may/09/schools.gcses2004.

Dennis, I. (2007) 'Halo effects in grading student projects'. *Journal of Applied Psychology*, 92 (4), 1169–76. https://doi.org/10.1177/0098628313487425.

Eduworld (2018) 'Researching motivators and drivers for international students and migrants when selecting a high-stakes test, to maximise client's marketing efforts'. Accessed 1 February 2020. www.eduworld.net.au/high-stakes-tests-case-study.

Falkner, N., Vivian, R., Piper, D. and Falker, K. (2014) 'Increasing the effectiveness of automated assessment by increasing marking granularity and feedback units'. *SIGCSE '14: Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, 9–14. https://doi.org/10.1145/2538862.2538896.

Foltz, P.W., Streeter, L.A., Lochbaum, K.E. and Landauer, T.K. (2013) 'Implementation and applications of the Intelligent Essay Assessor'. In M.D. Shermis and J. Burstein (eds), *Handbook of Automated Essay Evaluation*. New York: Routledge, 68–88. https://doi.org/10.4324/9780203122761.ch5.

Greatorex, J. and Bell, J.F. (2008) 'What makes AS marking reliable? An experiment with some stages from the standardisation process'. *Research Papers in Education*, 23 (3), 333–55. https://doi.org/10.1080/02671520701692593.

Isaacs, T. (2014) 'Curriculum and assessment reform gone wrong: The perfect storm of GCSE English'. *Curriculum Journal: Creating curricula: Aims, knowledge, and control*, 25 (1), 130–47. https://doi.org/10.1080/09585176.2013.876366.

James, M. (2006) 'Assessment, teaching and theories of learning'. In J. Gardner (ed.), *Assessment and Learning*. London: SAGE, 47–60. https://doi.org/10.13140/2.1.5090.8960.

Johnson, M., Hopkin, R., Shiell, H. and Bell, J.F. (2012) 'Extended essay marking on screen: Is examiner marking accuracy influenced by marking mode?'. *Educational Research and Evaluation*, 18 (2), 107–24. https://doi.org/10.1080/13803611.2012.659932.

Koreis, V. (no date) 'Čapek's R.U.R.'. Accessed 3 September 2020. https://web.archive.org/web/20131223084729/http://www.booksplendour.com.au/capek/rur.htm.

Liu, C.-L., Wang, C.-H. and Gao, Z.-M. (2005) 'Using lexical constraints to enhance the quality of computer-generated multiple-choice cloze items'. *Computational Linguistics and Chinese Language Processing*, 10 (3), 303–28. https://www.aclweb.org/anthology/O05-4001.pdf.

Luckin, R. (2017) 'Towards artificial intelligence-based assessment systems'. *Nature Human Behaviour*, 1. https://doi.org/10.1038/s41562-016-0028.

McDonagh, M. (2020) 'Why the exams debacle was so predictable – and predicted', *The Spectator*, 17 August. Accessed 25 September 2020. www.spectator.co.uk/article/when-it-comes-to-the-exams-fiasco-i-hate-to-say-i-told-you-so.

Moon, Y. and Pae, J.-K. (2011) 'Short-term effects of automated writing feedback and users' evaluation of criterion'. *Applied Linguistics*, 27 (4), 125–50. https://doi.org/10.1558/cj.26382.

Natarajan, T., Balasubramanian, S.A. and Kasilingam, D.L. (2017) 'Understanding the intention to use mobile shopping applications and its influence on price sensitivity'. *Journal of Retailing and Consumer Services*, 37, 8–22. https://doi.org/10.1016/j.jretconser.2017.02.010.

Newstead, S. and Dennis, I. (1994) 'Examiners examined: The reliability of exam marking in psychology'. *The Psychologist*, 7 (5), 216.

OECD (Organisation for Economic Co-operation and Development) (2020) 'PISA test: PISA, Programme for International Student Assessment (2018)'. Accessed 4 September 2020. www.oecd.org/pisa/test/.

Paschen, U., Pitt, C. and Kietzmann, J. (2020) 'Artificial intelligence: Building blocks and an innovation typology'. *Business Horizons*, 63 (2), 147–55. https://doi.org/10.1016/j.bushor.2019.10.004.

Pearson (2017) *Comparable Outcomes: A guide: How can improvements in teaching and/or learning be recognised under the comparable outcomes approach?* Accessed 7 December 2020. https://qualifications.pearson.com/content/dam/pdf/GCSE/support-documents/Comparable-Outcomes-a-Guide.pdf.

Pearson (2019) *Pearson Test of English Academic: Automated scoring*. Accessed 7 December 2020. https://pearsonpte.com/wp-content/uploads/2018/06/Pearson-Test-of-English-Academic-Automated-Scoring-White-Paper-May-2018.pdf.

Pearson Vue (2021) 'What we do'. Accessed 12 January 2021. www.pearsonvue.co.uk/About-Pearson-VUE/What-we-do.aspx.

Pinot de Moira, A. (2013) 'Features of a level-based mark scheme and their effect on marking reliability'. Centre of Education Research and Policy.

Popham, W.J. (2009) 'Assessment literacy for teachers: Faddish or fundamental?'. *Theory into Practice*, 48 (1), 4–11. https://doi.org/10.1080/00405840802577536.

Rhead, S. and Black, B. (2018) *Marking Consistency Metrics*. Coventry: Ofqual.

Richardson, H. (2020) 'Exam results: Where did it go wrong and what happens next?', BBC News, 17 August. Accessed 5 September 2020. www.bbc.co.uk/news/education-53811391.

Richardson, M., Baird, J.-A., Ridgway, J., Ripley, M., Shorrocks-Taylor, D. and Swan, M. (2002) 'Challenging minds? Students' perceptions of computer-based World Class Tests of problem solving'. *Computers in Human Behavior*, 18 (6), 633–49. https://doi.org/10.1016/S0747-5632(02)00021-3.

Richardson, M., Isaacs, T., Barnes, I., Swensson, C., Wilkinson, D. and Golding, J. (2020) *Trends in International Mathematics and Science Study (TIMSS) 2019: National report for England*. Accessed 12 January 2021. www.gov.uk/government/publications/trends-in-international-mathematics-and-science-study-2019-england.

Ripley, M. (2004) 'Expert Technologies Seminar on e-assessment: The e-assessment vision'. Presentation at BECTA Expert Technology Seminar. London: autumn.

Sadler, D.R. (1989) 'Formative assessment and the design of instructional systems'. *Instructional Science: An international journal of the learning sciences*, 18, 119–44. https://doi.org/10.1007/BF00117714.

Shaw, S. (2008) 'Essay marking on-screen: Implications for assessment validity'. *E-Learning*, 5 (3), 256–74. https://doi.org/10.2304/elea.2008.5.3.256.

Shermis, M.D. and Burstein, J. (2013) *Handbook of Automated Essay Evaluation: Current applications and new directions*. New York: Routledge. https://doi.org/10.4324/9780203122761.

Stewart, H. (2020) 'Boris Johnson blames "mutant algorithm" for exams fiasco', *The Guardian*, 26 August. Accessed 4 September 2020. www.theguardian.com/politics/2020/aug/26/boris-johnson-blames-mutant-algorithm-for-exams-fiasco.

Stiggins, R. (1995) 'Assessment literacy for the 21st century'. *Phi Delta Kappan*, 77 (3), 238–45.

Stobart, G. (2008) *Testing Times: The uses and abuses of assessment*. London: Routledge. https://doi.org/10.4324/9780203930502.

Stobart, G. (2009) 'Determining validity in national curriculum assessments'. *Educational Research*, 51 (2), 161–79. https://doi.org/10.1080/00131880902891305.

Timmis, S., Broadfoot, P., Sutherland, R. and Oldfield, A. (2016) 'Rethinking assessment in a digital age: Opportunities, challenges and risks'. *British Educational Research Journal*, 42 (3), 454–76. https://doi.org/10.1002/berj.3215.

Torney-Purta, J. and Amadeo, J.-A. (2013) 'International large-scale assessments: Challenges in reporting and potentials for secondary analysis'. *Research in Comparative and International Education*, 8 (3), 248–58. https://doi.org/10.2304/rcie.2013.8.3.248.

UK Parliament (2020a) 'Schools and colleges: Qualification results and full opening'. Hansard. Accessed 4 September 2020. https://hansard.parliament.uk/Commons/2020-09-01/debates/8CA678F8-D06B-4B21-B66B-8E3D4A448311/SchoolsAndCollegesQualificationResultsAndFullOpening.

UK Parliament (2020b) 'Ofqual questioned on summer exam results'. Accessed 7 December 2020. https://old.parliament.uk/business/committees/committees-a-z/commons-select/education-committee/news-parliament-2017/ofqual-evidence-19-21/.

Viera, A.J. and Garrett, J.M. (2005) 'Understanding interobserver agreement: The kappa statistic'. *Family Medicine*, 37 (5), 360–3.

Waring, M. and Evans, C. (2015) *Understanding Pedagogy: Developing a critical approach to teaching and learning*. London: Routledge.

Westfield, K. (2016) 'Students devastated A-Level exam papers STOLEN and exam board GUESSES their grades'. *Daily Express*, 18 August. Accessed 3 September 2020. www.express.co.uk/news/uk/701474/students-devastated-AQA-exam-board-loses-exam-papers-estimates-grades.

Whithaus, C. (2006) 'Always already: Automated essay scoring and grammar checkers in college writing courses'. In P.E. Ericsson and R. Haswell (eds), *Machine Scoring of Student Essays: Truth and consequences*. Logan: Utah State University Press, 166–76.

Wiliam, D. (2001) 'Reliability, validity, and all that jazz'. *Education 3–13*, 29 (3), 17–21. https://doi.org/10.1080/03004270185200311.

Young, S. (2001) 'Statistical modeling in continuous speech recognition (CSR)'. In J. Breese and D. Koller (eds), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, Seattle, WA, 2–5 August, 562–71. Accessed 12 January 2021. https://arxiv.org/abs/1301.4607v2.

Zawacki-Richter, O., Marín, V.I., Bond, M. and Gouverneur, F. (2019) 'Systematic review of research on artificial intelligence applications in higher education – where are the educators?'. *International Journal of Educational Technology in Higher Education*, 16, 39. https://doi.org/10.1186/s41239-019-0171-0.