The State-of-the-art in digital technology-based assessment

Abstract

The role of digital technology in assessment has received a great deal of attention in recent years. Naturally, technology offers many practical benefits, such as increased efficiency with regard to the design, implementation and scoring of existing assessments. More importantly, it also has the potential to have profound, transformative effects on the field of assessment by facilitating the integration of formative activities with accountability requirements and broadening the range of abilities and the scope of constructs that can be assessed. This article provides an overview of the current state-of-the-art in digital technology-based assessment, with particular reference to advances in the automated scoring of constructed responses, the assessment of complex 21st century skills in large-scale assessments and innovations involving high fidelity virtual reality simulations. Key challenges with respect to each are highlighted before the extent to which digital technology is truly transforming assessment is considered.

KEY WORDS

Assessment, collaborative problem-solving, digital technology, machine scoring, 21st century skills, virtual reality simulations

The state-of-the-art in digital technology-based assessment

Michael O'Leary, Darina Scully, Anastasios Karakolidis & Vasiliki Pitsia

Correspondence

Michael O'Leary, Centre for Assessment Research, Policy and Practice in Education (CARPE), St Patrick's Campus, Dublin City University, D09DY00 Dublin, Ireland
Email: michael.oleary@dcu.ie

Darina Scully, Centre for Assessment Research, Policy and Practice in Education (CARPE), St Patrick's Campus, Dublin City University, D09DY00 Dublin, Ireland
Email: darina.scully@dcu.ie

Anastasios Karakolidis, Centre for Assessment Research, Policy and Practice in Education (CARPE), St Patrick's Campus, Dublin City University, D09DY00 Dublin, Ireland
Email: anastasios.karakolidis@dcu.ie

Vasiliki Pitsia, Centre for Assessment Research, Policy and Practice in Education (CARPE), St Patrick's Campus, Dublin City University, D09DY00 Dublin, Ireland
Email:vasiliki.pitsia2@mail.dcu.ie

DISCLAIMER

The work of CARPE and the Chair in Assessment at DCU is supported financially by a grant from Prometric Inc– a test development , delivery and data management company headquartered in Baltimore, Maryland. Until recently, the beneficial owner of Prometric was the Educational Testing Service (ETS). The content of this paper has not been influenced in any way by either organisation, and is solely the responsibility of the authors.

1 INTRODUCTION

The interconnectedness of technology and assessment[1] has a long history. Madaus (2001) refers to a technology as any body of special knowledge, skills and procedures that people use to satisfy a need, solve a problem or attain a societal, economic, or educational goal. Under this definition, it is clear that educational assessment itself is a technology and, as Kellaghan & Madaus (2003) point out, one that has existed since an external civil service examination system was invented in China around 1100BC. While the system changed over the centuries until its demise in 1905, the general approach involved candidates writing up to eight essays within an allocated time period in which they explained ideas from the Four Books and Five Classics of Confucianism (www.sacu.org/examinations.html). In many respects, the artefacts and processes of assessment that make it a technology have been

visible from the very beginning in the form of test booklets and answer sheets, as well as in the specialist knowledge, skills and language of its community of practitioners. What has changed in the recent past is the speed with which one technology (digital technology[2]) is transforming another (assessment) in terms of the constructs that can be measured and the types of environments in which assessments can take place. This has not gone unnoticed by governments where the potential of digital technology to promote and measure the 21st century skills needed for economic prosperity, such as critical thinking and collaborative problem-solving, finds expression in a myriad of policy documents from around the world (OECD, 2016). Moreover, and as a consequence of over 20 years of research on how assessment for learning (AfL) can be used to enhance teaching and learning (Dumont, Istance, & Benavides, 2010), the affordances offered by digital technology in terms of how and when feedback is shared in the classroom are starting to be considered by a growing number of educational professionals and researchers. That said, the lack of alignment between most assessments in use and efforts to promote 21st century skills and other 'competence-based' approaches across curricula in Europe and elsewhere has also been noted (Adamson & Darling-Hammond, 2015).

The influence of digital technology on assessment is perhaps best understood with reference to three 'stages' of integration, as outlined by Bennett (2015). The first of these stages involves the delivery of traditional assessments via computers – a basic transition that for the most part takes limited advantage of technology. Second-stage technology-based assessment is characterised by incremental changes, including innovative item formats, the automation of various assessment processes and early attempts to improve the measurement of constructs (or aspects of constructs) that have proven difficult to measure using paper-and-pencil tests. The third stage is somewhat different. As Bennett (2015, p.372) explains, 'what was at first, an evolution driven primarily by technology becomes driven by substance'. At the superficial level, third-stage technology-based assessments may be identified by the fact that they often incorporate interactive performance tasks or simulations, or by their tendency to be 'more integrated with instruction, sampling performance repeatedly over time'. However, their most significant characteristic is that decisions about their design, content and format are informed by competency models[3] and by general cognitive principles from the science of learning. One such principle is that learning tends to be greatest when knowledge is contextualised or when new information is related to prior knowledge stored in memory (Dochy, 1992). Another is that individuals who are competent in a given domain tend to judge their performance against internalised standards about what constitutes 'good

performance' in that domain and that these internalised standards are more powerful than extrinsic standards enforced by others (Bandura, 1977). With these principles in mind, third generation assessments seek to situate problems in realistic contexts, tend to structure extended tasks in a progressive fashion and encourage learners to apply performance criteria to their work as part of the assessment process. In this sense, true third-generation technology-based assessments are essentially models of effective pedagogical practice (Bennett, 2015). They can be conceptualised not just as evolutionary, but as revolutionary.

In this article, we provide an overview of digital technology-based assessment across three broad domains of activity chosen to be illustrative of current state-of-the-art in the field. Specifically, we focus on (i) machine scoring of constructed responses with particular reference to automatic scoring of essays, (ii) innovations in large-scale assessments around collaborative problem-solving, and (iii) high fidelity virtual reality simulations that facilitate novel ways of presenting stimuli and gathering responses. As shall become evident throughout the article, the extent to which developments in each of these areas can be categorised as 'third-generation' assessments varies. Indeed, the barriers posed by the limits of current assessment technology are significant and will be important factors to consider in policy setting contexts for the foreseeable future.

2 ADVANCES IN MACHINE SCORING OF CONSTRUCTED RESPONSES

The history of machine scoring of tests dates back to the early 1930s when Reynold B. Johnson, a high school physics teacher in Michigan, began working on a device that could detect pencil marks on a sheet of paper using tiny electrical circuits. His subsequent employment at IBM led to the development of the company's 805 Test Scoring Machine and 'fill-in the bubble' type answer sheets. The first large-scale use of the IBM 805 was for the New York Regents exam in 1936 and it was launched commercially in 1937. While the electrical conductivity method was superseded by the more efficient optical mark recognition (OMR) technology in the 1960s, the basis of Johnson's innovation is still visible today in the automated scoring of millions of standardised tests and surveys worldwide[4]. Given this history, it is not surprising that the 1960s also saw Ellis Page begin his research to automate the process of evaluating and scoring constructed response items, although it was not until the 1990s when computing power was advanced enough that he developed a full working model of his *Project Essay Grade* or *PEG*® (Page, 1966, 1994). Others were also working on similar projects that led to the development of many of the machine scoring systems currently in use:

IntelliMetric® (Elliot, 2003), the *Bayesian Essay Test Scoring sYstem - BETSY*® (Rudnor & Liang, 2002), the *Intelligent Essay Assessor* ®(Landauer, Laham,& Foltz, 2003) and *e-rater*® (Attali & Burstein, 2006).

As Bejar, Mislevy and Zhang (2016) and Bennett (2015) note, at present, automated essay scoring (AES) is the most common application, although work that began in the last two decades to advance the automated scoring of speech (e.g., oral language proficiency, reading out loud), multimodal observations (e.g., classroom interactions, collaboration, interviewing skills, etc.) and content-based constructed-responses involving mathematical equations, medical simulations, architectural designs and the like continues unabated (Liu, Brew, Blackmore, Gerard, Madhok, & Linn, 2014; Loukina & Cahill, 2016). The remainder of this section will focus on AES and use the *e-rater* system to illustrate current state-of-the-art in the area before considering issues related to the technology more broadly.

e-rater is the automated scoring engine used at Educational Testing Service (ETS) to score the quality of writing in essays. Zhang (2013) points out that its use in the Graduate Management Admission Test (GMAT®) in 1999 made it the first automated scoring engine to be used in a high-stakes testing programme[5]. The *e-rater* engine uses natural language processing (NLP) and regression techniques to identify and weight features considered essential to high quality writing in order to predict the score a human rater would give to an essay. Chen, Zhang and Bejar (2017) describe it thus:

> … *e-rater* scores are generated by a linear combination of a set of high-level features computed for each essay with weights determined by regressing human ratings [from trained raters] on the features. These features are also called macrofeatures. Most of these macrofeatures are composed of sets of lower-level features called microfeatures that are combined to produce the macrofeature values. (pp.1-2)

They explain that in version 13.1 the macrofeatures used to predict human scores are organisation, development, grammar, usage, mechanics, style, word length, word choice, collocation and preposition and sentence variety, as well as two prompt-specific features relating to the content of vocabulary used when building a bespoke model for individual essays. They go on to explain that the value of a microfeature such as spelling is the count of the respective errors associated with it. The mechanics macrofeature, for example, is based on the total errors associated with 12 microfeatures related to spelling, capitalisation, punctuation, etc. The micro and macrofeature scores are then combined and weighted in a statistical model designed to produce an overall score that maximises agreement with human scoring. In essence, *e-rater* is 'trained' (programmed) to extract particular linguistic features

of writing in a model building process that is based on patterns observed in a large number essays representing different points on the entire scoring scale (e.g., a scale from 1 to 6). Specifically, the process requires a clearly-defined scoring guide with a matching essay topic, 350 responses per topic from the students to be assessed and scores from two trained independent readers using the scoring guide for each essay response. ETS explain that the weighting of the linguistic features (macro and micro) to assign a final score to an essay can be tailored to a particular prompt or can take a more generic approach whereby the same statistical model can be used to score responses to a variety of prompts (www.ets.org/erater/applications/). Ezzo and Bridgeman (2014) note that, whilst most automated scoring models are calibrated for a specific writing prompt, the developers of the *e-rater* engine, in emphasising form over content, take a more generic approach that improves score validity[6] by standardising linguistic features across prompts.

It is clear that in arriving at a final holistic summative 'judgement' through an analytic scoring process, *e-rater* is also capable of providing diagnostic feedback to test takers on the various linguistic features used in developing it. Indeed, in the context of research that indicates that feedback received beyond two weeks of a learning event tends not to have an effect (Wiliam, 2006), the fact that the *e-rater* engine has allowed ETS to develop web-based writing assessment services that are capable of providing students with real time detailed feedback on errors[7] made in a piece of writing on a topic is significant. The engine also provides an overall score corresponding to the one an instructor/faculty member would give, based on the scoring criteria for that topic (www.ets.org/criterion). The feedback provided through the extraction of the linguistic features (macro and micro) of a piece of written text is made possible by NLP, defined by Liddy (2001, p.1) as 'a theoretically motivated range of computational techniques for analysing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications'. It is an area of research that lies at the intersection of disciplines that include, but are not limited to, computational linguistics, mathematics, computer science, psychology and artificial intelligence (Chowdhury, 2003). According to ETS, *e-rater* is continually updated to reflect advances in NLP, whilst Chen et al. (2017) advocate for a critical and systematic review of the conceptual and statistical models at the heart of the engine. That said, the question of how scores derived from AES are used and the extent to which they provide a reliable[8] and valid alternative to human raters needs to be considered in a context beyond the specifics of how *e-rater* works.

Bejar et al. (2016) describe three approaches to using scores derived from current automated processes – a check score, a contributory score and a primary/sole score –, all of which can be applied to AES. In the first instance, AES provides a score that is used as a quality control mechanism to ensure that a human rater is sufficient. If a discrepancy beyond a threshold level occurs, a second or third human rater is used to determine the final score, but the AES score does not count. This approach is the one that is most likely to be used when the stakes associated with an essay are high, as they are for the GRE Analytical Writing assessment[9] (Bennett, 2015) or when cognitively complex skills are being assessed (Zhang, 2013). The second approach uses a weighted average of an AES score and a human score and, if a discrepancy results, using some combination of the AES score and two or more human scores. In this case, AES contributes to the final score. This method is used by ETS when scoring writing in the relatively low stakes TOEFL iBT[10] (Zhang, 2013). In the third approach, AES is used to determine a primary or sole score and human raters are only involved if the level of confidence in the AES score falls below some agreed threshold. Zhang (2013) notes that this approach is common in low stakes large-scale assessments[11] and in instances where the primary purpose of the writing is to provide the test taker with formative feedback.[12] Used in these three ways, AES has the potential to reduce measurement error by improving score reliability, to increase equity and fairness by providing access to more widely available and cheaper practice for high-stakes essay tests and to create multiple opportunities for feedback on learning, thinking, reading and writing (Landaur, Laham, & Folz, 2003).

It is well known that the human scoring of essays is not only labour intensive, expensive and time consuming, but also prone to validity and reliability problems caused by rater drift, halo effects, stereotyping, and so on. AES systems have developed to a point where they can be used to alleviate many of these problems. Their strength lies in the fact that they are efficient, impartial and objective, as well as being reliable in applying criteria and providing instantaneous feedback. They are also flexible enough to evaluate essays across grade levels and languages other than English (e.g., Intelligent Essay Assessor). However, Zhang (2013) points out that their current reality is that, in working with the features of writing that can be extracted and combined mathematically, more often than not they only evaluate 'relatively rudimentary text-production skills' (p.4)[13]. He goes on to argue that:

> Current automated essay-scoring systems cannot directly assess some of the more cognitively demanding aspects of writing proficiency, such as audience awareness, argumentation, critical thinking, and creativity. The current systems are also not well positioned to evaluate the specific content of an essay, including the factual correctness

> of a claim. Moreover, these systems can only superficially evaluate the rhetorical writing style of a test taker, while trained human raters can appreciate and evaluate rhetorical style on a deeper level.[14] p.4

Bennett (2015) expresses the same concern when referring to the 'elemental' nature of many of AES scored non-essay prompts that require clicking on hot spots or equation entry as responses. In 2003, Liddy made a point that is still relevant today. She explained that, in the early days of artificial intelligence research, the field of NLP was referred to as *Natural Language Understanding (NLU)* but that the term 'understanding' was dropped in favour of 'processing' because NLP systems cannot draw inferences from text. Today, NLU remains the ultimate goal of researchers working in the area.

As we approach the end of the 21st century's second decade, it is clear that the fields of artificial intelligence and NLP are not sufficiently advanced to allow AES systems to bypass the need to incorporate the scores from human ratings of sample essays in the programming process. A wealth of research indicates that, while automated and human scoring of essays are not exactly equivalent, the correlation between them in many instances is high, ranging in value from .60 to .96 (Ezzo & Bridgeman, 2014; Landaur et al., 2003; Shermis, 2014). However, this is a double-edged sword in validity terms. On the one hand, the research provides evidence to justify using AES as an alternative to human judgement scoring in certain circumstances, whereas on the other, it is premised on an assumption that the process of human judgement is well understood and provides a validity gold standard (Bridgeman, Trappani, & Attali, 2012). Unfortunately, the published research on human-rater cognition is sparse (Zhang, 2013) and the evidence demonstrating the superiority of human assessment has been lacking for some time (Bennett & Bejar, 1998; Bennett, 2015). In addition, Bennett and Zhang (2016) highlight the lack of sufficient research into six other dimensions of AES relative to human judgements in high stakes contexts: (i) the construct relevance of the scoring models used; (ii) quality assurance with respect to unusual responses involving, for example, atypical creativity or gaming/bluffing; (iii) the extent to which scores on one task generalise to scores on others within the same domain; (iv) the validity of assumed associations with measures of similar and different constructs; (v) the population invariance of scores (including gender, ethnicity and country comparisons – see Bridgeman et al., 2012); and (vi) the impact on teaching and learning practices.

All these lacunae notwithstanding, there is a clear sense that AES has brought many practical benefits to large-scale and local assessment initiatives, especially those in which the stakes for test takers are relatively low. In other contexts, however, the current reality is that

AES cannot be used as the sole arbiter of scores with a high degree of confidence. Liu et al (2014) reached the same conclusion regarding the automated scoring of concept-based constructed-response items using *c-rater*™, although report more positive findings (Liu, Rios, Heilman, Gerard, & Linn, 2016) with an updated version of the software called *c-rater-ML*™. Shermis (2015) also reports that the machine scoring algorithms produced outcomes that were not as consistent as human raters in the Hewlett Foundation-sponsored study involving nine automated systems. Given the ongoing and rapid rate of technological advances in so many spheres, one can expect that applications of AES will become more sophisticated, but the journey to the point where machines can understand the nuances of human communication may still be some time away.

3 DIGITAL TECHNOLOGY IN LARGE-SCALE ASSESSMENTS

Large-scale national and international comparative studies of student achievement are primarily designed to provide policy makers and others with good information about the functioning of educational systems (Johansson, 2016). Beginning in the 1960s and coming to prominence in the 1990s, the international studies are having an ever-increasing influence in a globalised educational world (Shute, Leighton, Jang, & Chu, 2016). Their sustained popularity and the need to remain relevant in terms of what is being measured and how have resulted in attempts to leverage technology to improve assessment practices and expand the range of what can be assessed. Beller (2013) explains that technology is now used in three ways in large-scale assessment contexts. First, it is used to facilitate the assessment of the domains that have traditionally been the focus in schools, namely reading, mathematics, and science. The main aim is to increase validity and improve the assessment of aspects within these domains that have previously proven difficult to assess. Secondly, technology is used to assess generic competencies, such as skills related to Information and Communication Technologies (ICT) and other transferable skills related to managing and communicating information. Thirdly, technology is used to assess more complex constructs, also known as 21st century skills. They include, but are not limited to creativity, critical thinking, learning to learn, entrepreneurship, problem-solving and collaboration (Beller, 2013; Riggio, 2014; Shute et al., 2016).

For some time now, computerised fixed-form tests have been used to create on-screen versions of traditional paper-and-pencil tests, whilst allowing for a greater range of response options, e.g., drag-and-drop, highlighting, selecting hot-spots, etc. (Redecker, 2013; Thompson & Weiss, 2009). Computerised variable-form tests, on the other hand, facilitate

the assessment of student knowledge and skills in ways that were not previously possible (Scheuermann & Bjornsson, 2009). Approaches here include adaptive testing[15] and linear-on-the-fly testing,[16] also known as automated test assembly in which the test items are configured during the examination time. In the last three decades, high-stakes adaptive tests have been mostly used for licensure and certification purposes, with early examples including the Novell corporation's certified network engineer (CNE) examination (1990), the Graduate Record Examination (GRE) (1992) and the National Council Licensure Examination-Registered Nurse in the U.S. (1994) (Lin, 2008; Luecht, 2005; Luecht & Sireci, 2012). In addition, according to Redecker (2013) and Shute et al. (2016), developments in digital-based task design (e.g., games and simulations) have not only enabled the assessment of multidimensional cognitive, metacognitive and affective learner characteristics in authentic contexts, but also, as Bryant (2017) and Redecker & Johannessen (2013) argued, have improved test-takers' engagement and motivation during the testing process. Finally, concerning skills such as problem-solving, technology use in international studies allows for the 'quiet assessment'[17] of the many aspects of authentic performance, such as students' intermediate products, strategies and thought processes (Bryant, 2017; Parshall, Harmes, Davey, & Pashley, 2010; Webb & Gibson, 2015).

The Organisation for Economic Co-operation and Development (OECD)'s Programme for International Student Assessment (PISA) has been at the cutting edge of large-scale test design, development, administration, data analysis and reporting since the first study was conducted in 1999. Of late, it has been at the forefront of efforts to measure individual and collaborative problem-solving skills.[18] PISA 2012 introduced a computer-based assessment of problem-solving skills where interaction between the problem solver and the problem was a major requisite for the solution of the problem (OECD, 2013). This work was extended, first with the development of a conceptual framework for collaborative problem-solving[19] and then with the administration of an assessment of collaborative problem-solving in PISA 2015 (OECD, 2017a; Webb & Gibson, 2015). Indeed, this innovation marked the first attempt in an international context to assess an element of students' social skills (OECD, 2017b).

Six units were designed for the purposes of the collaborative problem-solving assessment that were divided among the three clusters. Each comprised tasks/scenarios that required between five and 20 minutes to complete.[20] They were designed to measure three collaborative competencies (establishing and maintaining a shared understanding, taking appropriate action to solve the problem, and establishing and maintaining team organisation)

and four problem-solving competencies (exploring and understanding, representing and formulating, planning and executing, monitoring and reflecting). Collaborative problem-solving skills were measured through a wide range of items that each corresponded to one or more aspects of the collaborative and problem-solving processes (e.g., establishing and maintaining shared understanding and planning and executing) (OECD, 2017a).

The PISA 2015 assessment of collaborative problem-solving required each test taker to engage on screen with virtual computer agents in a range of collaborative processes. During the completion of the tasks, multiple measures of test takers' communications, actions, products and responses to probes were logged and categorised into three broad categories: group decision-making tasks, group co-ordination tasks and group-production tasks. Each produced a score that contributed towards one or more of the competencies. Although the computer agents could take on multiple roles and behaviours and test takers could make individual choices as they worked through the task, in reality, they were limited to those options that were already decided in the task development (e.g., statements corresponding to different levels of proficiency in a particular competency) and programmed into the software. This approach allowed for a high degree of control and standardisation, and, crucially in the context of a large-scale study, automatic scoring in most instances (OECD, 2017a). Additionally, all aspects of each individual's actions, communications, products and response times were saved as log files and processed through a fully-automated partial-credit scoring against each of the skills of the framework, using pattern-matching technology. This allowed for the identification of the key aspects of performance corresponding to the different competencies. In cases where constructed responses were required, a different offline scoring rubric was used to measure the quality of the communication and actions (OECD, 2017a).

The presentation of tasks on the screen and the students' inputs were based on conventional media and computer interface components, such as diagrams, figures, tables, interactive simulations, windows and icons, as well as actions involving mouse clicks, sliders (for manipulating quantitative scales), drag-and-drop, cut-and-paste and keystrokes. The interface also allowed for different communication methods, such as emails and menu-based chat interfaces (OECD, 2017a).

From the six collaborative problem-solving tasks designed for the study in 2015, only two were approved for public release and one of these, *The Visit*, will be used here as a state-of-the-art exemplar. Two moments from the PISA 2015 implementation of *The Visit* are captured in Figures 1 and 2. In the first part of the task, the test takers are presented with a

situation in which they are required to collaborate with three computer agent schoolmates (George, Rachel and Brad) to identify an appropriate local point of interest where a group of visiting international students can be taken. Specifically, the scenario requires the test takers to engage in a discussion about the essentials of a good visit and express a preference in relation to one of three possible locations. The test takers must also resolve an issue around the opening times and decide on a final plan. In all cases, the response options are limited to those already programmed into the task (see the lower left corner of Figure 1) (OECD, 2015).

Insert Figure 1 about here

In the second part, the test takers are presented with an email from the faculty adviser outlining additional requirements of the team relating to issues such as the ability to communicate in the visitor's native language, interests of the students and size of groups going on the visit. This information has been collected at the point presented in Figure 2 and the test takers and team members must then go on to address issues with respect to the amount of time being taken and how they could do better to meet the criteria next time (OECD, 2015).

Insert Figure 2 about here

Not surprisingly, given the focus on the assessment of higher-order skills using cutting edge technology, the PISA 2015 collaborative problem-solving initiative has garnered a good deal of attention worldwide and, as argued by He, von Davier, Greiff, Steinhauer & Borysewicz (2017), provided a major impetus for similar initiatives. For example, Fiore, Graesser, Greiff, Griffin, Gong, Kyllonen, and von Davier (2017), in considering an assessment framework for collaborative problem-solving in the National Assessment of Educational Progress (NAEP) in the US, use PISA 2015 to highlight the merits of using virtual agents to measure students' collaborative problem-solving skills[21]. However, they also make a case for the involvement of fewer virtual agents in the process, pointing out that involving two humans would lead to a more 'ill-defined' (less standardised) approach and more complex measurements.

More generally, the upgrade to computer-based testing in large-scale assessments is a relatively recent phenomenon and the focus has been predominantly on skills traditionally assessed using paper-and-pencil approaches and standard multiple-choice item formats (Pellegrino & Quellmaz, 2010; Redecker, 2013). Despite attempts to leverage the power of

digital technology to measure higher-order skills such as collaborative problem-solving, the resultant assessments are still quite structured and rigid, with complex forms of learning either being neglected or measured in ways that may not correspond to real-life situations. While current state-of-the-art technology can capture problem-solving processes such as strategies used, the nature and number of attempts and the time taken, it is not difficult to agree with those who argue that more robust and meaningful assessments of complex skills are still some way off (Beller, 2013; Bennett, 2015). As Webb & Gibson (2015) argue, whilst such ventures would obviously require the necessary financial investments, they also stress the need for time to be devoted to improving assessment literacy of those responsible for developing more robust assessments of complex skills.

4 INNOVATIONS IN ASSESSMENT USING VIRTUAL REALITY SIMULATIONS

Virtual reality (VR) has recently begun to feature in educational assessment. VR is a computer simulation of any 3D environment that creates a sense of being physically present in that environment (Linowes, 2015). Initially, cost prohibited its widespread use, but recent technological advances have led to something of a 'revolution', with high quality, user-friendly and more affordable applications now available. As Parisi (2015) noted, this development has been mostly driven by the gaming industry.

In the field of education, VR environments have been primarily used in instructional contexts, whilst examples of VR applications for assessment purposes are much scarcer (Redecker, 2013). Their effectiveness in improving learning outcomes has been evidenced by numerous studies, including a comprehensive meta-analysis exploring the benefits of using VR-based instructions for learning, spanning 69 studies from kindergarten to higher education (Merchant, Goetz, Cifuentes, Keeney-Kennicutt, & Davis, 2014). Most examples of VR used in learning activities stem from the field of medical education where it is employed to train students in practical tasks prior to their engagement with real patients. As Shute et al. (2016, p.51) argued, these 'technologically rich environments, the learning contexts in which many students find themselves, have created a need to reconsider quite dramatically the design and development of assessments'. Indeed, it makes little sense to use high quality VR to teach students how to conduct surgery and then administer multiple-choice tests to assess this heavily performance-based knowledge. Multiple choice tests, despite their reliability and objectivity, may not be sufficient indicators of 'real-life' performance (Agard & von Davier, 2018). VR environments, on the other hand, have the

potential to improve assessment fidelity and authenticity in that they allow learners to experience stimuli and perform tasks in ways that closely approximate real situations (Ryall, Judd, & Gordon, 2016).

With the use of VR environments, assessment can be transformed into an interactive process that goes beyond the measurement of knowledge recall and captures performance in complex tasks (Shute et al., 2016). At present, the most advanced applications of VR are in the field of personnel selection. However, given the extent of the reforms currently taking place in educational assessment and the emphasis on assessing complex skills and competencies in a more authentic manner (Darling-Hammond, Herman, Pellegrino, Abedi, Aber, Baker, & Steele, 2013) it is likely to be only a matter of time before these pioneering advances also infiltrate the field of education assessment.

In medical education, the equipment used in VR assessments typically consists of one or two levers and a monitor. The levers are used as simulators of surgical tools, with the more advanced of these providing haptic feedback to make the experience even more realistic, while the monitors provide 3D illustrations so that users can watch how the simulator interacts with their actions while performing the required tasks. Figure 3 presents an example of such a VR environment developed by MedaPhor®. It was initially designed for training purposes, however, Madsen, Konge, Nørgaard, Tabor, Ringsted, Klemmensen, and Tolsgaard (2014) investigated its use as an assessment instrument, and more specifically, its ability to distinguish novices (medical students) from experienced consultants. Their example is not the only one in the field. Ryall, Judd, & Gordon (2016) conducted a review of the literature on the use of various simulation-based techniques (e.g., standardised patients, anatomical

Insert Figure 3 about here

models, part-task trainers, computerised human patients and VR) in the assessment of technical skills in health education. It spanned 21 studies – a modest number that highlights the relative lack of simulation applications used specifically for assessment purposes. Only three focused on VR environments (Bick, DeMaria, Kennedy, Schwartz, Weiner, Levine, & Wagner, 2013; Grantcharov, Carstensen, & Schulze, 2005; Lipner, Messenger, Kangilaski, Baim, Holmes, Williams, & King, 2010). As Ryall et al. (2016) concluded, these studies, despite their small sample sizes, provided promising results in terms of the ability of VR assessment to distinguish between those who required further training and those who were ready for clinical practice.

Although there are few practical differences between the use of VR environments for training purposes and for assessment purposes, one key difference is the need to establish valid and reliable scoring procedures in the case of the latter. Some (Bick et al., 2013) have scored participants' performances in VR environments based on experienced raters' observations. However, this approach requires considerable resources and may not always provide reliable results. Madsen Konge, Nørgaard, Tabor, Ringsted, Klemmensen, and Tolsgaard (2014) on the other hand, adopted a computerised scoring approach which was initially developed by the VR manufacturers as a means of providing trainees with automated feedback. The VR simulator included various modules, ranging from basic to advanced, and upon completion of a module, automatically provided scores using dichotomous metrics (successful/unsuccessful) in a number of task-specific areas, as well as in general performance aspects. The final score for each participant was calculated by adding the scores (0 or 1) for each metric, an approach that showed high levels of internal consistency (Cronbach's alpha of 0.95)[22]. In terms of validity, the expert group performed better than the novice group, however, only one third of these metrics reliably distinguished students from experienced medical consultants and demonstrated evidence of construct validity. This indicates that dependency on metrics that have been developed to provide automated feedback during training may not always yield valid judgements of examinees' skills in formal assessment contexts. Despite their small scale, these examples suggest that such technologies have the potential to significantly improve practical assessments, not only in medical education, but also in other fields. However, more research is necessary.

It should be noted that in most of the available VR assessment applications, learners are seated and interact with the simulator via a computer screen. Although this kind of VR can offer high fidelity simulation in the case of certain tasks, such as that of ultrasound scanning, there are other situations in which it is more appropriate to afford the examinee the opportunity to interact with the virtual environment in more dynamic ways. Indeed, full VR technology enables 360-degree visual immersion through the use of cutting edge headset technology. Moreover, users can freely interact with the environment, using touch controller technology.

Figure 4 depicts a full VR assessment lab developed by recruitment company Capp®. This technology has already been piloted with a client of the company from the field of banking. One of the tasks involved in this assessment was 'The Responsible Business Challenge' – a competitive game where applicants were required to apply various skills in order to raise £250,000 for a children's charity. Rather than asking candidates to describe

how they would handle this situation, the VR technology enabled the recruiters to observe how people approached the task and how they dealt with any problems and challenges encountered.

Insert Figure 4 about here

VR assessments are appealing because they can capture multiple aspects of an individual's performance. Such technologies set new standards in what we can assess and how we assess it by providing the opportunity for performance skills (including the so-called '21st century skills') to be measured in standardised, yet authentic environments as opposed to static, question/response-based assessments.

As Almond, Kim, Velasquez, & Shute (2014) acknowledged, assessment in dynamic environments – whereby learners are required to perform tasks as opposed to answering a series of questions – is challenging. It is important to ensure that the tasks attempted by the learner and the ways in which various performances are scored accurately reflect the construct that the assessment purports to capture. In interactive assessments, such as VR simulations that afford learners considerable freedom to perform a task in a multitude of ways, it is extremely difficult to predict all outcomes. Of course, those designing these assessments have the option to limit the space of the possible behaviours in the VR environment, but this may impact on its fidelity. All these concerns render the measurement element of VR assessments extremely complex (Levy, 2012). As a general rule, VR applications that were originally designed solely for instruction purposes may not be easily converted into assessment 'instruments'. This is the case because, as Mislevy, Behrens, Dicerbo, Frezzo, & West (2012) argued, careful consideration of scoring strategies should be included as part of the process of designing a VR assessment and not arise only when this developmental stage is complete. Indeed, developing logical scoring that takes advantage of the plethora of information in VR assessments is probably the main challenge to be addressed in this field.

5 DISCUSSION

Viewing digital technology-based assessment through Bennett's (2015) lens, it should be recognised that many of the activities described in this article are best categorised in the second stage of integration. AES, for example, has the potential to enhance the field of assessment by increasing the efficiency of an existing practice, but it falls short of

transforming assessment in terms of facilitating the measurement of complex competencies or re-conceptualising the principles that guide assessment design. Moreover, the validity of AES currently rests on the questionable assumption that the judgement of human raters on which it is based is infallible. Initial attempts to integrate technology into international large-scale assessments, such as the computer-based assessment of science literacy in PISA 2006, are also illustrative of the second stage of development. By incorporating a variety of innovative item types intended to reduce reading/writing demands and increase student motivation, these assessments were an important step in terms of harnessing technology to improve the validity of judgements about students' proficiency in an existing construct. It is only more recently, however, that these international studies – and indeed assessment in various other contexts – have begun to enter the third stage of technology-based assessment.

Both the PISA 2015 assessment of collaborative problem-solving and the use of virtual reality technology in medical education bear characteristics of Bennett's 'third generation'. That is, the design of each of these assessments has been driven both by the need to assess complex constructs (e.g., social skills involved in problem-solving, competency to perform clinical procedures) in an authentic manner and by key cognitive principles of learning (e.g., that learning can be enhanced when knowledge is contextualised). In each of these examples, third-generation goals are beginning to be realised through the development of sophisticated interactive environments or intelligently-designed tasks that elicit and encourage the expression of complex constructs and processes. However, many challenges still remain and it is important to be cognizant of the limitations associated with any new development in technology-based assessment. As discussed previously, although high fidelity simulations extend the range of possible behaviours that can be exhibited during an assessment, this presents difficulties for the development of reliable and valid scoring procedures that can take all these possibilities into account. Similarly, incorporating digital technology into international assessments raises questions, such as (i) how to detect meaningful trends with previous administrations, and (ii) how to make valid comparative inferences about education systems and cultures that vary in terms of their readiness to engage with technology, and in how competencies such as collaborative problem-solving are understood.

It is clear that the field of assessment is undergoing great changes with the influence of digital technology. From a practical viewpoint, technology has improved the efficiency of many aspects of assessment delivery and scoring; and more recently, in parallel with advances in computing and artificial intelligence, it has opened up possibilities for

increasingly complex, sophisticated and intellectually challenging assessments. That said, it is also clear that we are only on the cusp of realising its full potential. As Adamson and Darling-Hammond (2015) noted, references to 21st century skills are now firmly established in curricular frameworks and policy documents worldwide, but in reality, these skills are heterogeneous, and practical efforts to assess them still lag behind. This is particularly true in the case of less cognitively-oriented skills, such as citizenship and personal and social responsibility. In order to ensure that future developments in technology-enhanced assessment take positive steps towards narrowing this gap, it is important to critically evaluate the contribution of each new innovation. Ultimately, those involved in assessment design would do well to bear in mind Bennett's description of third generation technology-based assessment as 'driven by substance'. It is imperative that technology does not become the primary focus of 21st century assessment, with the emphasis remaining on reliability, validity, authenticity and underlying pedagogical purpose.

REFERENCES

Adamson, F. & Darling-Hammond, L. (2015). Policy pathways for twenty-first century skills. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach*. Dordrecht: Springer Science and Business Media.

Agard, C. & von Davier, A. (2018). The virtual world and reality of testing: Building virtual assessments. In H. Jiao & R. W. Lissitz (Eds.), *Technology enhanced innovative assessment: Development, modeling and scoring from an interdisciplinary perspective*. Charlotte, NC: Information Age Publishing.

Almond, R. G., Kim, Y. J., Velasquez, G., & Shute, V. J. (2014). How task features impact evidence from assessments embedded in simulations and games. *Measurement*, *12*, 1-33.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater® v. 2.*The Journal of Technology, Learning, and Assessment, 4*, 2-29.

Bandura, A. (1977). Self-reinforcement: The power of positive personal control. In P. G. Zimbardo & F.L. Ruch (Eds.), *Psychology and life (9th ed.).* Gelnview, IL: Scott, Foresman.

Beller, M. (2013). Technologies in large-scale assessments: New directions, challenges, and

opportunities. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research*. Dordrecht, NL: Springer.

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education, 39*, 370-407.

Bennett, R. E., & Behar, I. I. (1998). Validity and automated scoring: It's not only in the scoring. *Educational Measurement: Issues and Practice,17*, 9-17.

Bennett, R. E.& Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement*. New York, NY: Taylor & Francis.

Bejar, I, Misleavy, R, & Zhang, M. (2016). Automated scoring with validity in mind. In A. A. Rupp & J.P. Leighton (Eds.), *The handbook of cognition and assessment. Frameworks, methodologies and applications*. Hoboken, N.J. Wiley Blackwell.

Bick, J. S., DeMaria, S., Kennedy, J. D., Schwartz, A. D., Weiner, M. M., Levine, A. I., & Wagner, C. E. (2013). Comparison of expert and novice performance of a simulated transesophageal echocardiography examination. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare, 8*, 329-334.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*, 27-40.

Bryant, W. (2017). Developing a strategy for using technology-enhanced items in large-scale standardised tests. *Practical Assessment, Research and Evaluation*, *22* (1).

Capp. (n.d.). *Real assessment in a virtual world.* Retrieved from http://capp.co/virtualreality

Chen, J., Zhang, M., & Bejar, I.J. (2017). An investigation of the e-rater® automated scoring engine's grammar, usage, mechanics, and style: Microfeatures and their aggregation model. (Research Report No. RR-17-04). Princeton, NJ: Educational Testing Service.

Chowdhury, G. (2003) Natural language processing. *Annual Review of Information Science and Technology, 37*, 51-89.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education.* (7th ed.). London, England: Routledge.

Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aber, J. L., Baker, E., & Steele, C. M. (2013). Criteria for high quality assessment. *Stafford Center for Opportunity in Policy in Education*, 25. Retrieved from http://edpolicy.stanford.edu/publications/pubs/847

Dumont, H., Istance, D., & Benavides, F. (Eds.) (2010). The nature of learning: Using research to inspire practice. Paris: OECD Publishing.

Dochy, F.J.R.C. (1992). *Assessment of prior knowledge as a determinant for future learning: The use of knowledge state tests and knowledge profiles.* Utrecht/London: Lemma/Jessica Kingley Publishers.

Elliot, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum Associates.

Ezzo & Bridgeman (2014). Overview of automated scoring for the GRE® General Test. In C Wendler & B. Bridgeman (Eds.), *The research foundation for the GRE® revised general test: A compendium of studies (*Section 4.1.1 - 4.1.5). Princeton NJ: Educational Testing Service.

Fiore, S. M., Graesser, A., Greiff, S., Griffin, P., Gong, B., Kyllonen, P., & von Davier, A. (2017). *Collaborative problem solving: Considerations for the National Assessment of Educational Progress.* Retrieved from https://nces.ed.gov/nationsreportcard/pdf/researchcenter/collaborative_problem_solving*.pdf*

Grantcharov, T. P., Carstensen, L., & Schulze, S. (2005). Objective assessment of gastrointestinal endoscopy skills using a virtual reality simulator. *Journal of the Society of Laparoendoscopic Surgeons*, *9*, 130-3.

Griffin, P. & Care, E (2015). *Assessment and teaching of 21st century skills: Methods and approach*. Dordrecht: Springer Science and Business Media.

He, Q., von Davier, M., Greiff, S., Steinhauer, E. W., & Borysewicz, P. B. (2017). Collaborative problem solving measures in the Programme for International Student Assessment (PISA). In A. von Davier, M. Zhu, & P. Kyllonen (Eds.), *Innovative assessment of collaboration*. Cham, Switzerland: Springer International Publishing.

Johansson, S. (2016). International large-scale assessments: What uses, what consequences? *Educational Research*, *58*, 139-148.

Kellaghan T. & Madaus G. (2003) External (Public) examinations. In T. Kellaghan & D.L. Stufflebeam (Eds.), *International handbook of educational evaluation*. Dordrecht, NL: Kluwer Academic Publishers.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum.

Landaur, T.K., Laham, D., & Folz, P. (2003). Automated essay scoring. *Assessment in Education, 10*, 295-308.

Levy, R. (2012). *Psychometric Advances , Opportunities, and Challenges for Simulation-Based Assessment*. Center for K-12 Assessment and Performance Management Report, Educational Testing Service.

Liddy, E.D. (2003). Natural Language Processing. In M.A. Drake (Ed.). *Encyclopedia of Library and Information Science, 2nd Ed.* New York, NY. Marcel Dekker, Inc. Retrieved from: https://surface.syr.edu/cgi/viewcontent.cgi?referer=http://scholar.google.com/&httpsredir=1&article=1043&context=istpub.

Lin, C.J. (2008). Comparisons between classical test theory and item response theory in automated assembly of parallel test forms in automated assembly of parallel test forms. *The Journal of Technology, Learning and Assessment*, *6*, 4-42.

Linowes, J. (2015). *Unity Virtual Reality: Explore the World of Virtual Reality by Building Immersive and Fun VR Projects Using Unity 3D*. Birmingham, England: Packt Publishing.

Lipner, R. S., Messenger, J. C., Kangilaski, R., Baim, D. S., Holmes, D. R., Williams, D. O., & King, S. B. (2010). A technical and cognitive skills evaluation of performance in interventional cardiology procedures using medical simulation. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*, *5*, 65-74.

Liu, L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M.C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice, 33*, 19-28.

Liu, L., Rios, J.A., Heilman, M., Gerard, L., & Linn, M.C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching 53*, 215-233.

Loukina, A., & Cahill, A. (2016). *Automated scoring across different modalities*. Downloaded from: www.aclweb.org/anthology/W16-0514.

Luecht, R. M. (2005). Some useful cost-benefit criteria for evaluating computer-based test delivery models and systems. *Journal of Applied Testing Technology, 7*. Retrieved from www.jattjournal.com/index.php/atp/article/view/48338

Luecht, R. M., & Sireci, S. G. (2012). *A review of models for computer-based testing.* Retrieved from https://files.eric.ed.gov/fulltext/ED562580.pdf.

Madaus, G. (2001). *Educational testing as a technology. National Board on Educational Testing and Public Policy Statements*, 2(1). Retrieved from www.bc.edu/research/nbetpp/publications/v2n1.html

Madsen, M. E., Konge, L., Nørgaard, L. N., Tabor, A., Ringsted, C., Klemmensen, A. K., & Tolsgaard, M. G. (2014). Assessment of performance measures and learning curves for use of a virtual-reality ultrasound simulator in transvaginal ultrasound examination. *Ultrasound in Obstetrics and Gynecology, 44*, 693–699.

MedaPhor. (n.d.). *SCANTRAINER: Transvaginal simulator*. Retrieved from www.medaphor.com/scantrainer/transvaginal-simulator/

Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., & Davis, T. J. (2014). Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis. *Computers and Education, 70,* 29-40.

Mislevy, R., Behrens, J., Dicerbo, K., Frezzo, D., & West, P. (2012). Games, learning, and assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning*. New York, NY: Springer.

OECD (2013). *PISA 2012 Assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris, France: OECD Publishing.

OECD (2015). *PISA 2015 Released Field Trial Cognitive Items* (ETS, Ed.).

OECD (2016). *Skills for a digital world: Background Paper for Ministerial Panel 4.2 - DSTI/ICCP/IIS(2015)10/FINAL*. Paris: OECD/Directorate for Science, Technology, and Innovation/Committee on Digital Economy Policy/Working Party on Measurement and Analysis of the Digital Economy.

OECD (2017a). *PISA 2015 Assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving (revised edition).* Paris, France: OECD Publishing.

OECD (2017b). *PISA 2015 Results (Volume V): Collaborative problem solving*. Paris, France: OECD Publishing.

Page, E. B. (1966). The imminence of grading essays by computer, *Phi Delta Kappan*, *48*, 238-243.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62,* 127-142.

Parisi, T. (2015). *Learning virtual reality: Developing immersive experiences and applications for desktop, web, and mobile*. Sebastopol, CA: O'Reilly Media.

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. (2010). Innovative items for computerized testing. In W. J. van den Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (2nd ed.). Norwell, MA: Kluwer Academic Publishers.

Pellegrino, J. W. & Quellmaz, E. S. (2010). Perspectives on the integration of technology and assessment. *Perspectives on the Integration of Technology and Assessment, 43,* 119-134.

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

Redecker, C. (2013). *The use of ICT for the assessment of key competences*. Luxembourg: European Commission: Joint Research Centre, Institute for Prospective Technological Studies: European Union. https://doi.org/10.2791/87007

Redecker, C. & Johannessen, O. (2013). Changing assessment--Towards a new assessment paradigm using ICT. *European Journal of Education, 48*, 79-96.

Riggio, R. (2014). *The "hard" science of studying and developing leader "soft" skills*. *Leader interpersonal and influence skills*. New York, NY Routledge.

Rudnor, L.M. & Liang, T.(2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning, and Assessment, 1*, 3-21.

Ryall, T., Judd, B., & Gordon, C. J. (2016). Simulation-based assessments in health professional education: A systematic review. *Journal of Multidisciplinary Healthcare, 9*, 69-82.

Scheuermann, F. & Bjornsson, J. (Eds.) (2009). *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing*. *EUR 23679 EN - 2009*. Luxembourg: European Commission: Joint Research Centre, Institute for the Protection and Security of the Citizen: European Communities.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition results and future directions from a United States demonstration. *Assessing Writing, 20,* 53-76.

Shermis, M. D. (2015). Contrasting state-of-the-art in the machine scoring of short-form constructed responses. *Educational Assessment, 20*, 46-65.

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art in machine scoring of essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation*. New York, NY: Routledge.

Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M.-W. (2016). Advances in the science of assessment. *Educational Assessment, 21*, 34-59.

Thompson, N.A & Weiss, D.J. (2009). Computerized and adaptive testing in educational assessment. In F. Schueuermann & J. Bjornsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale assessment (EUR 23679 EN- 2009).* Luxemboug: European Commission: Joint Research Centre, Institute for the Protection and Security of the Citizen: European Communities.

Webb, M. & Gibson, D. (2015). Technology enhanced assessment in complex collaborative settings. *Education and Information Technolgies, 20*, 675-695.

Wiliam, D. (2006). Formative assessment: Getting the focus right. *Educational Assessment, 11*, 283-289.

Zhang, M. (2013*). Contrasting automated and human essay scoring of essays. (R&D Connections No. 21).* Princeton, NJ: Educational Testing Service.

NOTES

[1] Assessment is used here as an umbrella term to include testing, measurement and all aspects of assessment processes that involve gathering, recording, using and communicating information about learning.

[2] The terms digital technology and Information and Communication Technology (ICT) are used synonymously.

[3] Competency models emphasise the need to assess the full range of a target domain – i.e., the integration of knowledge and skills in the performance of a function – as opposed to mere factual knowledge.

[4] Details are taken from www-03.ibm.com/ibm/history/ibm100/us/en/icons/testscore/

[5] A high-stakes assessment/test is one that has important consequences for the person or entity being assessed. For example, any assessment that determines who obtains a college place can be considered high stakes.

[6] According to the latest *Standards for Education and Psychological Testing*, validity can be defined as *"the degree to which evidence and theory support the interpretations of test scores for purposed uses of test"* (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 2014, p. 11).

[7] *e-rater* provides feedback on grammar, word usage, mechanics, style, organization, and development. This is also a feature of other AES systems such as Intelligent Essay Assessor™ which provides feedback on six aspects of writing — ideas, organization, conventions, sentence fluency, word choice, and voice (Zhang, 2013).

[8] Broadly speaking, reliability is linked to precision and accuracy (Cohen, Manion, & Morrison, 2011). More specifically, it refers to the consistency of scores across replications of a testing process (AERA, APA, & NCME, 2014).

[9] The Graduate Record Examination (GRE) consists of a suite of tests developed by ETS and used by universities in the US and elsewhere as one of the criteria for entry to various graduate programmes.

[10] The *TOEFL iBT* is a test measuring candidates' ability to use and understand English at the university level.

[11] Zhang cites ETS's *TOEFL*® Practice Online (TPO), the College Board's ACCUPLACER®, and ACT's COMPASS® as examples.

[12] Zhang cites *MY Access*® (Vantage Learning), *WriteToLearn*® (Pearson Education, Inc.), and the *Criterion*® Online Writing Evaluation Service (ETS) as current state-of-the-art technologies that can do this.

[13] Following an extensive study of nine AES systems used in the K-12 context in the US (see, Shermis & Hammer, 2013, Shermis, 2014), Perelman (2014) responded with an article entitled: When "state of the art" is counting words".

[14] A useful summary of strengths and weaknesses in human and AES is provided in Zhang (2013), p. 5.

[15] An adaptive test tailors the difficulty of an item to the examinee's ability, as determined by their responses to the previous items (Luecht & Sireci, 2012).

[16] In linear-on-the-fly testing, a computer program pseudo-randomly selects items such that all examinees are presented with tests that are equivalent, but composed of entirely different items (Luecht, 2005).

[17] 'Quiet assessment' (also called stealth assessment) is an unobtrusive approach to the measurement of competencies, where examinees are not aware of being assessed (Shute et al., 2016).

[18] The focus on problem solving, particularly collaborative problem solving in PISA is not surprising given the arguments (empirically and philosophically based) for its importance to future learning and effective participation in globalised economies (e.g. Scheuermann & Bjornsson, 2009).

[19] Other initiatives that influenced the PISA framework include: the Assessment and Teaching of 21st-Century Skills (ATC21s), the Programme for the International Assessment of Adult Competencies (PIAAC) and the Partnership for 21st-Century Skills (OECD, 2017a).

[20] Each student was assigned one two-hour test form composed of four 30-minute "clusters. Two clusters were devoted to science, the major domain, and the remaining time was assigned to either one or two of the additional domains - reading, mathematics and collaborative problem solving - on the basis of a rotated test design (OECD, 2017a).

[21] The authors also review the work undertaken as part of the ATC21S initiative in Australia. Like PISA, ATC21S project designed tasks to elicit collaborative problem solving behaviours but involved students working in pairs and communicating through on-screen chat messaging while solving game-like puzzles. The test was adaptive in the sense that the difficulty of the tasks was adjusted based on various parameters such as the complexity of the items and skills needed to solve the problem. All data were saved in log files and following coding, calibrations of the data were undertaken for collaborative problem solving as well as social and cognitive skills, participation, perspective-taking, social regulation, task regulation and knowledge building. The ATC21S human-to-human approach offered higher levels of face validity in so far as interactions during problem solving were more realistic. However, the fact that it provided a less standardised assessment environment meant that scoring was problematic (Fiore et al., 2017; Griffin & Care, 2015).

[22] Cronbach's alpha values that are greater than 0.7 are accepted as an indication of sufficient internal consistency in a measure (Cohen et al., 2011).