2013

Using digital technologies to improve the authenticity of performance assessment for high-stakes purposes.

Christopher P. Newhouse
Edith Cowan University, p.newhouse@ecu.edu.au

10.1080/1475939X.2013.851031 This article was originally published as: Newhouse, C. P. (2013). Using digital technologies to improve the authenticity of performance assessment for high-stakes purposes. Technology, Pedagogy and Education. Available online at here

Using digital technologies to improve the authenticity of performance assessment for high-stakes purposes

This paper reports on the outcomes of a three-year study investigating the use of digital technologies to increase the authenticity of high-stakes summative assessment in four Western Australian senior secondary courses. The study involved 82 teachers and 1015 students and a range of digital forms of assessment using computer-based exams, digital portfolios, and audio-visual recordings. The results were analysed using a feasibility framework concerning manageability, technical facility, functional operation and pedagogic alignment. By the end of the study each form of assessment that was implemented was found to be feasible once some obstacles were overcome. Two methods of marking were tried, analytical rubric-based marking and holistic comparative pairs marking with the latter found to generate more reliable scores. With the increased use of digital technologies in schools and the expectation that children will achieve more complex performances more use of digital forms of assessment will be required.

Keywords: summative assessment, authentic assessment, comparative pairs, computer-based exam, digital portfolio, digital video

Introduction

Students tend to be more engaged when learning includes practical tasks (e.g., Ridgway, McCusker, & Pead, 2006) but teachers are often more concerned with accountability for test scores (Clarke-Midura & Dede, 2010), and thus as President Barak Obama puts it, tend to "teach to the test" and make "education boring for kids" (eSchool News, 2011, p. 15). It is right that teachers should be accountable for the investment a society makes in education, but societies such as in Australia (MCEETYA., 2008), now expect that children should not just gain theoretical knowledge but should be capable of applying knowledge for practical performance in life and workplace situations, as embodied in Singapore's "Teach less, learn more" policy (Ministry of Education, 2009). At the same time it is becoming clear that complex performance enables students to gain deeper conceptual knowledge (Clarke-Midura & Dede, 2010). It therefore follows that assessment of learning should address complex practical performance (Pellegrino & Quellmalz, 2011) and as

McGaw (2006) explains this needs to be taken seriously by education authorities even when the stakes are high.

> If tests designed to measure key learning in schools ignore some key areas because they are harder to measure and attention to those areas by teachers and schools is then reduced, then those responsible for the tests bear some responsibility for that. (p. 3)

Traditionally in Western Australia (W.A.) high-stakes assessment has largely been conducted using paper and pen exams. Research tends to indicate that this form of assessment would be inadequate for measuring complex practical performances or deep conceptual knowledge and thinking, such as required for problem-solving (Clarke-Midura & Dede, 2010; Lane, 2004; Lin & Dwyer, 2006; Stobart & Eggen, 2012). The problem is that the form of assessment lacks construct validity and as McGaw (2006) points out when the stakes are high there is also a concern for consequential validity, that ironically when linked with concerns about reliability has been an argument against using digital technologies (Clarke-Midura & Dede, 2010). Further, the lack of validity appears obvious in courses where students learn with and about digital technologies but are assessed using the old paper-based technologies (Clarke-Midura & Dede, 2010; Ridgway, et al., 2006).

Alternatives to paper-based exams have been, and are used, to a varying extent for high-stakes purposes in jurisdictions such as W.A. They have tended to rely on human observation, interview, physical portfolio submission and even audiovisual recording for evidence of performance. However, their use has been limited by the cost, logistical complexity, and difficulty in obtaining reliable measures of performance (Lin & Dwyer, 2006). McGaw (2006) argues that many of these limitations may be overcome through the combined use of digital technologies and modern psychometric methods. Taylor (2005) suggests that the use of digital technology for student assessment is “inevitable” (p. 9). However, Dede (2003) cautions that there are " fundamental barriers to employing these technologies”, mainly “psychological, organizational, political and cultural" (p.9), that Hall (2010) argues ensures a complex and slow process of implementation, a

'bridge'. This underscores the need for research to demonstrate the feasibility of digital forms of assessment, and that this delivers easier and higher quality results (Dede, 2008).

This paper summarises and generalises from the results of a three-year study that investigated the potential to use digital technologies to support more authentic forms of practical performance assessment in four senior secondary courses in W.A.: Applied Information Technology (AIT), Engineering Studies, Italian Studies, and Physical Education Studies (PES). The study was conducted by the Centre for Schooling and Learning Technologies (CSaLT) at Edith Cowan University (ECU) in collaboration with the Curriculum Council of Western Australia and supported by an Australian Research Council (ARC) Linkage research grant. As the intention of this paper is to provide an overview of, and conclusions from, the entire project it does not include much of the fine detail, particularly of the data collected, and does not report separately on the school case studies. Some of this detail has been previously published (Newhouse, 2010, 2011).

Design and Development of Digital Assessments

The major aim of the study was to increase the validity of summative assessment in the courses by leveraging the affordances of digital technologies (Stobart & Eggen, 2012). Dochy (2009, p. 105) discusses the manner in which "new assessment modes" may improve the validity of the tasks, the scoring, generalisability, and consequential validity. He explains that initially construct validity "judges how well assessment matches the content and cognitive specifications of the construct being measured". In the study this was achieved using course teams, a situation analysis, and seeking the perceptions of teachers and students. If this is done then Dochy claims the authenticity and "complex problem characteristics" of the task improves its validity. Secondly he explains that criteria to judge student performances need to be fair and allow demonstration of ability. In the study this was addressed through the use of standards-referenced analytical marking and holistic comparative pairs marking, and through correlation analyses between these methods of marking. Thirdly, Dochy explains how generalisability can be improved through a consideration of reliability. In the study this was addressed through a combination of Rasch model analysis, and

inter-rater correlation analysis.

Irrespective of the form of assessment a student performance needs to be judged by assessors according to some criteria (Stobart & Eggen, 2012). Therefore assessors need to either observe the performance, as in a music recital, or a representation of the performance, such as an art exhibition. Largely due to cost and logistics high-stakes assessment has tended to require assessors to judge representations, mainly on paper. It is now possible to use a range of digital representations including audiovisual recording and student created representations using appropriate software (Clarke-Midura & Dede, 2010). These representations can be created relatively easily and inexpensively and provided for assessors to judge using digital repositories and computer networks.

The form of assessment and design of the tasks that lead to digital representations of performance are critical to the functional quality of the assessment (Pellegrino & Quellmalz, 2011). Based on a review of the literature four digital forms of assessment were defined in the study: an *Oral Recording*, an *Extended Production Exam*, a *Focussed Performance Tasks Exam* and a *Reflective Digital Portfolio*. Sadler (2009) and Dochy (2009) provide longer lists of common forms appropriate for the assessment of complex performances. An *Oral Recording* may occur using a stand-alone video camera or a web-cam or microphone connected to a computer (e.g., Shrosbree, 2008). An *Extended Production Exam* addresses a practical problem with a full set of processes to develop a solution. An example that the study built upon extensively was the eScape project by Kimbell et al (2007) applying a design process. A *Focussed Performance Tasks Exam* comprises a series of tasks, not necessarily logically connected, to demonstrate particular knowledge and skills (e.g., Boyle, 2006). A *Reflective Process Digital Portfolio* was an organised collection of annotated digital artefacts. Digital portfolios provide a range of authentic ways of assessing practical performance manageably (Ridgway, et al., 2006; Taylor, 2005).

The Study

Most of the courses in senior secondary school in W.A. have practical components and there is an

expectation from students and the community that the assessment of student performance will reflect the nature of this learning. However, typically summative assessment is predominantly by paper and pen despite references to practical performance and higher order thinking skills in the course syllabi. A small number of courses include other forms of assessment such as interviews, recitals or physical portfolios but typically these are expensive and logistically complex, particularly for highly dispersed student cohorts. Therefore the study set out to investigate the feasibility of using digital forms of assessment to replace paper and pen or other non-digital alternatives to increase authenticity and/or cost-effectiveness. Feasibility was defined using the eScape project framework (Kimbell & Wheeler, 2005) in terms of manageability, technical facility, functional operation (i.e. validity and reliability) and pedagogic alignment. The main research question was: How are digitally based representations of student work output on authentic tasks most effectively used to support highly reliable summative assessments of student performances for courses with a substantial practical component? This question needed to be interpreted for each of the four courses.

Research Problem for each Course

For the *AIT* course, in contrast to the other three courses in the study, digital technologies provided the content for study as well as pedagogical support. Therefore performance related to using the technologies to demonstrate capability in using the technologies. The syllabus states that the course "provides opportunities for students to develop knowledge and skills relevant to the use of ICT to meet everyday challenges". It was intended that the majority of time be spent in using digital technologies to develop information solutions. It should therefore be surprising that the external assessment consisted of a three-hour paper-based exam. There were a number of ways in which digital technologies could have been used for performance assessment, principally either forms of portfolio or computer-based exam. The research question therefore became, which form of assessment was more feasible for the course?

The *Engineering Studies* course was a new course structured with a design core, and one of three specialization areas. The external assessment consisted of a 3-hour written examination that measured student knowledge on both the core and specialization areas. For a practical performance based subject the exam did not reflect that essential nature of the course. Consequently pedagogies were too theoretical and teachers had difficulty connecting theory and practice. The assessment of practical performance was considered too logistically challenging and expensive and thus the question concerned the potential for using digital technologies to support the assessment of the application of knowledge to practical work in a cost-effective and manageable form.

The *Italian Studies* course had a tradition of assessing oral performance through a face-to-face 'interview' at a central location in Perth, where two markers assessed each student's performance in real time. This resulted in logistical difficulties for students and the organising body and raised questions of reliability with the real-time scoring and lack of an enduring record of the process. Therefore the question was whether digital recordings of oral performance could be as adequately equivalent to the face-to-face 'interview' for the purposes of summative assessment, while delivering cost and increased reliability benefits.

The *PES* course was a new course and thus an external assessment had to be designed and implemented that would "celebrate the physical in physical education" (Penney, Evans, & Taggart, 2005, p. 7) and clearly promote integration of conceptual and performance-based learning. In 2008 the external assessment comprised a written and a practical exam that required students to attend centres located at sporting facilities in Perth to complete sporting skills and strategic response scenarios. Performances were video recorded and subsequently marked but this was discontinued because the resource requirements were not sustainable. There were also concerns about the rigour and authenticity of the assessment because students did not reveal the intentions of their response, adapt responses to changing circumstances, apply theoretical knowledge to a given situation, or evaluate their performance. Therefore the research question became how could these aspects be added to the practical assessment in a cost-effective manner?

Method

The study involved teacher-class case studies for the four courses with a total of 82 teachers and 1015 students involved. For the first two years the samples of case studies was selected to ensure teachers were likely to be able to implement the assessment tasks, while in the third year the aim was to get a more representative sample from the W.A. schools systems. For each course a common assessment task was developed that typically consisted of a number of sub-tasks. Student performance was captured in different ways in the four courses, for example, by audiovisual recording, typing responses and creating digital products. All the captured evidence of performance was stored in digital files on central servers.

Two methods of marking, analytical standards-referenced and holistic comparative pairs, were used to judge student performance based on the digital evidence. Analytical marking involved two or three assessors using sets of criteria, represented in rubrics, generated from the assessment tasks and the curriculum syllabi. Comparative pairs marking involved a larger number of assessors making comparisons between pairs of performances as explained by Pollitt (2012). Judgments were guided by a holistic criteria based on the analytical criteria. For example, for the Engineering Studies exam the criteria was, – 'The student is able to progress from an initial idea, in response to a range of stimulus and activities, to a satisfactory solution in a manner that clearly communicates the rationale for doing so'. The reliability of the data from both methods of marking was investigated using the traditional Cronbach's Alpha statistic through Rasch polytomous and dichotomous modelling respectively. Correlation analysis was used to both investigate reliability and compare the resulting scores from the different methods of marking.

All marking was done using online tools accessible through a standard web browser. In the first year online marking tools were built using *FileMaker Pro* for each method of marking for each course. These tools typically presented the digital representation of student performance on one side of the screen and recorded the judgements of assessors on the other. These judgements could then be exported to spreadsheets for analysis. In subsequent years comparative pairs marking was

implemented using and extending the *Adaptive Comparative Judgement System (ACJS)* that is associated with the MAPS portfolio system (Pollitt, 2012). This system included the analysis of assessor judgements.

In addition to the scores from marking a variety of qualitative data was collected by observing students working on the assessment tasks, surveying the students on completion or the tasks, and interviewing groups of students, all teachers and assessors.

Findings and Generalisations

The study initially aimed to provide recommendations on changes to assessment in the four courses but also intended to generalise findings beyond these courses to be relevant to other jurisdictions and other discipline areas. This section provides a summary of findings of the study for each course and then seeks to generalise these findings using the feasibility framework.

For each course the intention was to implement the same assessment task for each case although there was scope for some local customisation for the AIT portfolio and the PES task was customised for the sporting context. This intention was achieved by the third year for all four courses where, in each case, implementation of the assessment task was facilitated by either a researcher or trained invigilator with assistance from the teacher. In the first two years for the AIT and Italian Studies portfolios, which the classroom teacher largely facilitated, there were variations on the extent to which the assessment task was implemented as intended. Table 1 provides a summary of the inter-rater correlations for analytical marking (a measure of reliability) and the correlations between the scores from analytical and comparative pairs marking. For all courses in all years a Cronbach's Alpha reliability coefficient of at least 0.90 was achieved for comparative pairs marking while for analytical marking this was achieved by the third year.

<TABLE 1 HERE>

Applied Information Technology

In general terms the performance to assess in AIT was a student's response to a challenge given in

the form of a design brief leading to the development of a prototype digital product. Thus the end product was necessarily captured in digital form. However, the design and development processes also needed to be captured digitally. Therefore initially the assessment task was a four-week digital portfolio consisting of three components (a product, a process document, and two extra artefacts), and a three-hour computer-based production exam. For the third year the portfolio was discontinued. The main differences between these two forms of assessment were the time available and thus the scope of the challenge and set of processes enacted. In the exam students only had two or three hours, had no opportunity to investigate the challenge, were guided through the design and development processes and had more limited access to digitising tools with, for example, no access to cameras and the Internet. The study found that in typical schools both the portfolio and the exam could be adequately implemented to assess a range of levels of performance.

For the portfolio a sample design brief was supplied but teachers were encouraged to create a replacement relevant to their students. Each year the exam had a different design brief but with a similar structure, and with all digital resources provided (Figure 1). In response to recommendations by the teachers, the design brief was made increasingly more open-ended and provided increasing choices of product types. Each year assessors, teachers and students acknowledged the assessment tasks to be faithful to the course syllabus. The portfolio was facilitated by the teacher during normal class time, but the exam was facilitated by a researcher and the teacher, and delivered on a USB flash drive.

<FIGURE 1 HERE>

During the first two years there were 13 classes involved while in the third year 17 classes completed the two-hour production exam. The intention was that teachers would embed the assessment task within their teaching programmes and assessment structures. However, often this did not occur with it forming, in whole or in part, an additional task. The portfolio was implemented with some variation between teachers although all adhered to the requirements more rigorously in the second year. Most teachers used the sample design brief that led to the development of a web

site for a store as the product that provided adequate scope for students to demonstrate their capability. The quality of the process document varied considerably. The two extra digital artefacts provided scope for presentation of a broader range of skills that was not realised for most students. Unfortunately many submitted two websites or two slideshows that did not demonstrate a breadth of skill.

Each year the exam was successfully implemented for all classes with the main difficulties in a few schools being access to a computer laboratory, and activation of either USB ports or sound cards. The type of product created in the last two years appeared to align with student capability with for example, lower capability students creating posters compared with higher capability students creating websites or animations. Most students appeared comfortable with the initial design sub-task (most using paper that was then scanned), reflecting on their design and development processes at the end (typed into a template document), and including graphics, video and sound within their prototype product, but not with spreadsheets.

Analytical marking resulted in a good spread of scores for both the portfolio and exam with moderate to high significant correlations between the analytical assessors (Table 1). However, the correlations between the scores on the portfolio and exam were generally low to moderate indicating differences in the performances that they were measuring. The scores from analytical marking of the exam were not as reliable as those for the portfolio (α=0.85 cf 0.94). Rasch analysis showed that very few top scores were given on a few criteria and thus with minor modifications to the scoring adequately reliable scores were generated (α=0.90). These scores were at least significantly moderately correlated with the scores from comparative pairs marking. However, the two methods generated substantial differences in the ranking of a number of students, particularly evident in the final year.

Students and teachers tended to be drawn to the practical components of their course and perceived that the portfolio and computer-based exam were appropriate for assessing those components. Almost all preferred practical assessment to paper-based theory exams, although less

so for final year students. Students and teachers tended to indicate greater familiarity with the portfolio although the strict structure was less familiar. Teachers liked the flexibility of the portfolio but recognised the greater ease of invigilating the exam. At least three referred to limitations of their infrastructure including difficulties in working with audio, problems with servers and networks, and some unreliable hardware. A few students had concerns about malfunctions and a lack of time during computer-based exams. However, for the few students for whom there was a malfunction they were able to continue at an alternative workstation.

Each year the assessors tended to refer to the quality of work being lower than expected partly as a result of the relatively low level of sub-tasks required in the exam. This was to ensure universal access to appropriate software and skills to attempt the tasks. This was not a problem with the portfolio with the design brief being relatively open-ended and permitting customisation by the class teacher. Overall there was a balance between allowing students to demonstrate understanding and providing adequate scaffolding. When students were permitted greater choice and freedom that appeared to allow higher achieving students to demonstrate their capability. There was a danger of stifling the opportunity for students to demonstrate understanding through over-structuring tasks for marking convenience.

Engineering Studies

Over the three years there were 11 schools, 14 teachers and 21 classes of secondary students involved. The focus of the assessment task was on the core of the course and was a production exam with a scaffolded series of design iteration sub-tasks from a problem scenario, outlined in a design brief (e.g. to design a solar shower or a means of collecting water on a desert island), to a final solution. Successive iterations of the design followed some form of stimulus input, leading to students revising their ideas that were then reflected in sketches, models and audiovisual recordings. In the first two years the design development culminated in the creation of a 3D physical model and students had three hours that was reduced to two hours in the third year with the removal of the physical modelling. A digital design portfolio emerged through the input of text, graphics,

photographs, audio or video in response to the sub-tasks. In the first year the task was presented, and the digital output collected, using a custom-built *Filemaker Pro* database system running from a USB flash drive. In the final two years the *e-scape* exam management system was used (Kimbell & Wheeler, 2005).

The e-scape system provided three implementation methods, the 'intranet', 'live', or 'USB drive', with the appearance to students being almost identical. The first method was via the school intranet using a set of netbook computers, wirelessly linked with the facilitator's laptop. The portfolios automatically uploaded to the facilitator's computer and were later manually uploaded to the *e-scape* server in the UK. The second method was live via the Internet with the students logging on directly to the *e-scape* server. For the first two methods student progress could be controlled and monitored from the facilitator's computer. The final method utilized a 'runtime' version of the system on a USB flash drive for each student. Students had to progress at their own rate with the drives collected and the files later manually uploaded to the server.

In the first two years of the study the analytical assessors gave a different range of scores with considerably different means and standard deviations and only a moderate to weak correlation between them (Table 1). In the third year there was a significant moderate correlation between the two external assessors' scores indicating reasonable reliability and appropriate marking guides. In the third year Rasch polytomous modelling found that three criteria were considered to require modification due to disordered thresholds. Rescoring yielded a more reliable result (α=0.89). Overall the comparative pairs method generated a more reliable set of scores than the analytical method. Correlation analysis found a significant moderate relationship between the scores generated by the two methods of marking in the first year but only low to moderate correlation in the subsequent years. This probably indicated that the two methods of marking were assessing different aspects of student performance with the analytical marking perhaps more focussed on technical details.

Over the three years it was shown that a computer-based production exam could be used to assess core content of the course and be delivered in a variety of ways to suit the digital infrastructure available in the schools. The cyclic design process scaffolded four iterations of design that perhaps could be reduced if students were experienced in using the system and were provided with more comprehensive initial information. The two-hour time for exam appeared to be adequate, as indicated by the students who paced themselves through the exam using the USB drives. Students indicated that using a computer made all of the sub-tasks easier and that they preferred practical work to a focus on theory. Generally they indicated that they believed that the exam did provide an adequate opportunity to demonstrate their skills and understanding although some students, and also some teachers, did not consider that the task represented an engineering problem.

Italian Studies

The main performance focus of the study for the Italian Studies course was on oral communication. Initially the assessment revolved around audiovisual recording of oral language responses to stimuli. Over the three years a number of approaches were taken with the ultimate aim to simulate a conversation using digital technologies that could be accommodated within typical schools.

The first year involved Year 11 students in four schools with the assessment task consisting of a portfolio of sub-tasks leading to a video-recorded oral presentation. However, only students from one school submitted complete portfolios, but students from all the schools made an oral presentation that was video-recorded. Generally the students were extremely nervous about being video recorded that was likely to have inhibited their performance. However, many perceived that the digital technologies enabled them to reflect critically on their performance that would help them in preparing for future oral assessments. The scores from analytical and comparative pairs marking of the presentations were found to be highly reliable ($\alpha>0.90$) and there was a moderate to strong correlation between them (Table 1).

In the second year the aim was to align as closely as possible with the oral ‘interview’ exam that was the existing external practical assessment for the course. Therefore Year 12 students from

six schools were involved in completing an assessment task that consisted of an in-class computer-based exam (a series of oral recordings to visual stimuli) to replace the portfolio and a video recorded interview that mirrored the 'interview' exam. The in-class computer-based exam was delivered on USB flash drives using the 'runtime' version of the *e-scape* system. Preparing the drives and subsequent uploading of each student's recordings was a time-consuming task and clearly not readily scalable. Firewall issues had prevented the online version of the system being used. Students appreciated being able to logon to the system to access feedback provided by an assessor. The recorded interviews were completed with no technical issues and with the two assessors, the student, and a digital video camera in the room.

Teachers and students felt that the computer-based sub-tasks were of an appropriate standard. However, the teachers believed that the tasks did not accurately reflect a conversation and would be better if modified so that the students listened to, rather than read, the stimulus questions. In general, the teachers did not believe that a computer-based task could simulate a conversation although teachers and students recognised that the tasks were useful preparation for the recorded interview. They did believe that the video recording of the interview could lead to a fairer assessment process overall. Scores from analytical marking of the two forms of assessment were found to be highly reliable (α>0.90) and there was a reasonably strong correlation between them (r=0.74). This was an indication that the computer-based recordings were measuring the same construct as the interview and thus may possibly be used to replace the interview in the high stakes assessment. These scores were also highly correlated (Table 1) to the comparative pairs scores for the recorded interviews.

In the third year the assessment task involved Year 11 students completing an online exam with two components - Listening and Responding, and Oral Communication. For the first component students listened to an audio program in Italian and typed answers to questions. The aim of the second component was to simulate a conversation by combining aspects of both components from the previous year and thus three short videos were produced in-house using actors. Students

watched each video and made an audio recording for each in response to a question posed by one of the actors. Both components were delivered through *Willock Information System's* online testing system on their server in Canada. Unfortunately this system proved to be unreliable in uploading the audio recordings for the second component with up to 50% failure at some school sites. Further, in two schools with MacBooks the audio recordings could not be made.

For the analytical marking the work was accessed through this online system using its marking module. For comparative pairs marking recordings were first downloaded and then uploaded by a research assistant into the MAPS system. The analysis of the scores from both methods of marking showed a high level of reliability (α=0.89 to 0.90). The assessment task as a whole was acknowledged to be faithful to the course and it successfully distinguished between students with various levels of mastery. Both students and teachers found the Listening and Responding task to be at least equal to the traditional form of assessment for this outcome. However, the Oral Communication component was considered to be inferior to the traditional face-to-face 'interview' exam due to the technical problems and the perceived nature of simulating a conversation.

In the third year it was determined that the affordances provided by the oral recording in the computer-based exam outweighed the limitations. The major benefits were that students could proceed at their own pace, critically reflect upon their performance, and have a distraction-free listening experience using headphones. In general schools had the required technologies to undertake these digital forms of assessments. The study demonstrated that the use of computer-based oral recording has promise as a high-stakes form of oral performance assessment in place of face-to-face interview, delivering logistical benefits to both the students and the awarding body.

Physical Education Studies

The aim for the PES course was to enhance the practical examination that focuses on students' ability to make decisions and apply skills to resolve tactical challenges encountered in a sporting context. The digital assessment aimed to address issues of sustainability and authenticity while also

ensuring rigour. Thus strategies were developed to capture the intentions of students in their strategic responses, require them to adapt responses to changing circumstances, reflect on or evaluate their performance, and to apply theoretical knowledge, concepts and principles. The structure of the task was designed to allow customisation for sporting contexts used in PES (e.g. volleyball, netball, soccer, rugby, tennis, cricket and swimming).

Initially a four-part performance tasks exam was developed that centred on a tactical challenge in the sport the student was studying. The first part involved using a computer to respond to questions about the challenge, how to address it, and what skills and/or movements would be required. In the first year responses could only be typed but thereafter a drawing tool was added. The second part was video-recorded and required the demonstration of four skills related to the challenge. The third part was also video-recorded and put students into 'limited game' situations that required them to respond to the challenge and an adaptation of the challenge. The fourth part involved students using computers to view their video clips and respond to reflective and evaluative questions about their performance. Over the three years apart from minor changes to wording, the assessment task had very few changes. For the third year groups of students were assessed at the same time in the modified 'game' scenarios rather than one at a time. This typically required the use of coloured numbered bibs to aid in identifying students in the videos. Also after the first year wherever possible the first three parts were conducted contiguously.

A portable remote-controlled multiple camera system was used that allowed up to four cameras mounted on tripods to be controlled by a joystick (zoom, tilt and rotation) and feed into a single file on a laptop computer. The cameras were each connected by a single cable that provided the power, control and video feeds that were displayed as either quarter or full screen images.

Over the three years, there were 11 schools, 18 teachers and 19 classes of students involved. All four parts of the performance task exam were completed successfully with all groups of students with varying technical and logistical obstacles to overcome. Generally, students and teachers perceived the assessment as relevant to the course and preferred to a written exam. It was also

acknowledged as not dissimilar to the learning experiences that teachers were endeavouring to build into their teaching. Teachers considered that the task had achieved a degree of connection between traditionally distinct 'theoretical' and 'practical' components of PES. Most teachers tended to indicate that the task was superior to the current practical exam conducted in the course.

The assessors had some technical difficulties accessing the videos with both the analytical and pairs marking tools. Some videos were difficult to see and follow individual students and for analytical marking the videos were not streamed. There was some degradation of video quality with the *ACJS*. Over the three years improvements in quality were made and identification of students was enhanced using coloured bibs. Despite the difficulties, assessors were able to make appropriate judgements and input these using the marking tools from many locations. There were concerns about the reliability of the scores from analytical marking, despite a person separation index of 0.96, due to a low inter-assessor correlation (Table 1), and Rasch analysis identifying model misfit particularly for two criteria. However, using the comparative pairs method generated reliable scores (α=0.95). There was a significant moderate to strong correlation between the analytical and comparative pairs scores. There were substantial differences in the ranking of the performances of some students by the different methods of marking. This probably suggests that the holistic judgements made with the comparative pairs method was more suited than the analytical standards-referenced method.

Generalising the Findings using the Feasibility Framework

The study used a four-dimensioned feasibility framework to investigate the effectiveness of each form of assessment. Conclusions are summarised below using these dimensions with reference to the different types of digital forms of assessment.

Manageability Dimension

This dimension concerned the manageability of the assessment task in school environments (e.g. logistics of facilities, equipment and students) and providing the resulting output for assessors (e.g.

preparing and uploading digital files). For production or performance tasks exams there were few logistical difficulties except where a computer laboratory had to be specially booked. Invigilators needed specific knowledge and skills and spacious rooms with at least 10% excess computers to allow relocation in the case of technical problems. Practice sessions were needed to familiarise students. There were some logistical difficulties in managing portfolios, most concerning ensuring students organised their time to complete all requirements and adhere to defined conditions.

The time and expertise required for the management of the resulting digital files varied depending on the types of technologies used. This involved tasks such as reformatting and/or renaming files, checking the integrity of the files, and uploading them to online repositories. The most substantial management was required for the AIT tasks because students created the files and used a large range of software. Much less management was required where digital systems and/or facilitators created a limited set of files, as was the case in PES where the *Filemaker Pro* database system and the video recording system created all the files. Once the files were uploaded to the online repositories the marking processes required very little management with online marking offering additional affordances such as access from a variety of locations, efficient storage and backup, and sharing of data. In all situations the use of the *ACJS* considerably improved the manageability of comparative pairs marking.

Technical Dimension

This dimension concerned the adequacy of the performance of the technology. In all cases except for a few schools in Engineering Studies, school computing infrastructure was used. The main technical difficulties arising were connection to external servers through the Internet, and in some cases audio recordings and USB connections. Any online system needed to be adaptable to a variety of technical provisions such as firewall restrictions and network capacities across the variety of hardware, browsers, and operating systems. In some cases, despite extensive testing (especially under load) this was not possible, particularly for uploading of audio or video recorded responses. Therefore systems such as eScape that had provision for local storage in addition to any attempted

streaming to a server minimised data loss. The multi-camera remote control video-capture system used in PES proved to be reliable and flexible including support for underwater filming, provision of feedback to facilitator through an iPad, and control by a single operator. The use of digital cameras, web-cams and microphones for the other courses resulted in very few technical issues.

Functional Dimension

The functionality of the assessments implemented is discussed in terms of validity and reliability.

Overall it could be concluded that for each course the assessment task had adequate validity and certainly more so than the alternative of a paper and pen exam. This was aided by the use of a situation analysis to guide the development of assessment tasks. Digital forms of assessment enabled a variety of types of student responses (e.g. written, drawing, audio, video). In general the students and teachers perceived the digital forms of assessment used to be authentic, meaningful and contributing to connecting the theoretical and practical components of the courses. Generally they preferred this to the existing external assessment except for in Italian Studies where a face-to-face interview was preferred. Each form of assessment could be structured to permit a good range of levels of achievement to be demonstrated although this was more difficult for AIT.

For each course the assessment task eventually could be judged to provide scores with adequate reliability. Analytical marking using rubrics constructed specifically for the task generally resulted in at least moderately reliable scores. Consistently the comparative pairs method of marking was found to yield reliable scores. However, typically significant differences in ranking were found between the two methods of marking. The extent of the discrepancy most often depended on the extent to which the task was seen as holistic in nature with the comparative pairs method generally better suited to more open-ended performance-based tasks with a single main outcome. Where there were a number of diverse outcomes an analytical approach was best.

Pedagogic Dimension

This dimension considers the extent to which assessment tasks align with the pedagogies employed

for a course. Typically, students liked doing the practical work, appreciated the opportunity to demonstrate their creative capability, and were happy to respond where they could photograph, type and draw, less so for audio responses. Although in general, teachers viewed positively the match between the digital form of assessment and the intended pedagogy for a course there was evidence that many did not regularly implement this pedagogy (e.g. AIT and Engineering Studies) and this tended to be reflected in the quality of student performance captured. In all courses some aspects of the assessment were foreign to students and therefore they needed to be prepared with practice tasks to ensure they could perform optimally. Where courses such as Italian Studies have a well-established set of pedagogies tied to particular historic and traditional assessment models it is difficult to change these using digital forms of assessment.

Conclusion

The study showed a variety of ways in which the affordances of digital technology could feasibly deliver more authentic forms of assessment despite some constraints. The affordances varied for the different forms of assessment. Reflective portfolios were appropriate when a range of knowledge and skills needed to be demonstrated with a comprehensive set of processes. Production exams allowed the demonstration of a limited set of knowledge and skills to some depth of but within a limited set of processes. Performance tasks exams were best suited to addressing a range of knowledge and skills aside with little regard for sets of processes. The main constraints were logistical in organising time to complete the tasks and in some cases technical in either running software on school workstations or accessing online systems through school networks. As new technologies emerge this is an area for further research to ensure that using the technology delivers a better outcome (Dede, 2008).

The affordances of digital technology for judging, or marking, were clearly demonstrated for both methods, analytical and comparative pairs. The online database systems used allowed ready access for assessors from anywhere with adequate Internet bandwidth. The storage, backup and transmission of digital representations of student performances and assessor judgements were

easily achieved. The adaptive comparative judgements method of marking typically generated more reliable scores than analytical marking, but is probably only valid when the assessment task is fundamentally holistic (i.e., not made up of many required sub-tasks) with a minimum of scaffolding. This method of marking is relatively untested for high-stakes assessment (Pollitt, 2012) and therefore our findings need replication and extension.

Generally students are highly amenable to digital forms of assessment preferring them to paper-based forms provided that they have confidence in the hardware and some experience with the software and form of assessment. Even for Italian Studies, where students were less positive about the assessment, they perceived value in improving their performance. Almost all students are able to quickly learn how to use simple types of software required in assessment. Teachers are amenable to digital forms of assessment provided that the benefits to students are clear, implementation is relatively simple, and any software is easy to learn. However, as Hall (2010) notes moving from acceptance of the idea to implementing it, is not trivial. We found that experienced teachers and graduate students could be trained to implement digital forms of assessment but this needs testing for each variation.

In jurisdictions such as Western Australia, almost all secondary schools have large investments in digital technologies aimed to improve the relevance and engagement of learning for students. At the same time curriculum both locally and nationally is intended to deliver the high level of knowledge and skills, and their application, required for 21st Century societies. Such curriculum will only be adequately implemented if it includes highly authentic forms of assessment that this study has demonstrated can be supported by leveraging the digital technologies available in schools (Clarke-Midura & Dede, 2010). There seems little purpose in developing a 21st Century curriculum and investing in digital technologies in schools if that technology is not employed to support more authentic forms of performance assessment.

Acknowledgement

The study discussed in this paper was the work of a research team led by Paul Newhouse and John Williams and included senior researchers Dawn Penney, Cher Ping Lim, Jeremy Pagram, Andrew Jones, Martin Cooper, Alistair Campbell, and research assistants and postgraduate students. The work of everyone in this team, and the teachers involved, has contributed to the research outcomes presented in this paper.

References

Boyle, A. (2006). An evaluation of the decision to base the key stage 3 ICT test on a bespoke virtual desktop environment. London, U.K.: Qualifications and Curriculum Authority.

Clarke-Midura, J., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research on Technology in Education, 42*(3), 309-328.

Dede, C. (2003). No cliche left behind: why education policy is not like the movies. *Educational Technology, 43*(2), 5-10.

Dede, C. (2008). Theoretical perspectives influencing the use of Information Technology in teaching and learning. In B. K. J. Voogt (Ed.), *International Handbook of Information Technology in Primary and Secondary Education* (Vol. 20, pp. 43-62). New York: Springer Science + Business Media, LLC.

Dochy, F. (2009). The Edumetric Quality of New Modes of Assessment: Some Issues and Prospects. In G. Joughin (Ed.), *Assessment, Learning and Judgement in Higher Education* (pp. 85-114). Wollongong: University of Wollongong.

eSchool News. (2011). Obama: Too much testing makes education boring. *eSchool News, 14*(5), 15.

Hall, G. E. (2010). Technology's achilles heel: achieving high-quality implementation. *Journal of Research on Technology in Education, 42*(3), 231-253.

Kimbell, R., & Wheeler, T. (2005). Project e-scape: Phase 1 Report. London: Technology Education Research Unit, Goldsmiths College.

Kimbell, R., Wheeler, T., Miller, A., & Pollitt, A. (2007). e-scape: e-solutions for creative assessment in portfolio environments. London: Technology Education Research Unit, Goldsmiths College.

Lane, S. (2004). Validity of High-Stakes Assessment: Are Students Engaged in Complex Thinking? *Educational Measurement, Issues and Practice, 23*(3), 6-14.

Lin, H., & Dwyer, F. (2006). The fingertip effects of computer-based assessment in education. *TechTrends, 50*(6), 27-31.

MCEETYA. (2008). *Melbourne Declaration on Educational Goals for Young Australians*. Melbourne: Education Services Australia Limited Retrieved from http://www.curriculum.edu.au/verve/_resources/National_Declaration_on_the_Educational_Goals_for_Young_Australians.pdf.

McGaw, B. (2006). *Assessment to fit for purpose.* In conference proceedings, 32nd Annual Conference of the International Association for Educational Assessment, Singapore, pp.

Ministry of Education. (2009, 20th May 2009). Teach less, learn more. Retrieved 8th March, 2013, from http://www3.moe.edu.sg/bluesky/tllm.htm

Newhouse, C. P. (2010). Aligning Assessment with Curriculum and Pedagogy in Applied Information Technology. *Australian Educational Computing, 24*(2), 2-5.

Newhouse, C. P. (2011). Using IT to assess IT: Towards greater authenticity in summative performance assessment. *Computers and Education, 56*(2), 388-402. doi: 10.1016/j.compedu.2010.08.023

Pellegrino, J. W., & Quellmalz, E. S. (2011). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education, 43*(2).

Penney, D., Evans, J., & Taggart, J. (2005). *Educational codes and inclusivity in Physical Education.* In P. L. Jeffery (Ed.), conference proceedings Creative Dissent: Constructive Solutions, Australian Association of Research in Education Conference, Sydney, Australia., pp. 177-187.

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281-300.

Ridgway, J., McCusker, S., & Pead, D. (2006). Report 10: Literature Review of E-assessment. In K. Facer (Ed.), *Futurelab Series*. Bristol, UK: Futurelab.

Sadler, D. R. (2009). Transforming Holistic Assessment and Grading into a Vehicle for Complex Learning. In G. Joughin (Ed.), *Assessment, Learning and Judgement in Higher Education* (pp. 45-64). Wollongong: University of Wollongong.

Shrosbree, M. (2008). Digital video in the language classroom. *JALTCALL Journal, 4*(1), 75-84.

Stobart, G., & Eggen, T. (2012). High-stakes testing - value, fairness and consequences. *Assessment in Education: Principles, Policy & Practice, 19*(1), 1-6.

Taylor, A. R. (2005). A Future in the Process of Arrival: Using Computer Technologies for the Assessment of Student Learning. (pp. 119). Kelowna, British Columbia: Society for the Advancement of Excellence in Education.

Table 1. Correlations between scores generated by the two methods of marking and between analytical assessors.

		Correlation Coefficients					
		Between Methods of Marking			Inter-rater for Analytical Marking		
Course	Form	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
AIT	Production Exam	0.73**	0.77**	0.68**	0.70**	0.67**	0.69**
	Portfolio	-	-	-	0.90**	0.84**	-
Engineering Studies	Production Exam	0.78**	0.30**	0.46**	0.43**	0.26*	0.53**
Italian Studies	Audio Recording	0.70**	0.84**	0.77**	0.93**	0.77**	0.88**
PES	Performance Exam	0.89**	0.69**	0.73**	0.87**	0.49**	0.61**

** Correlation is significant at the 0.01 level (2-tailed). * Correlation is significant at the 0.05 level (2-tailed).

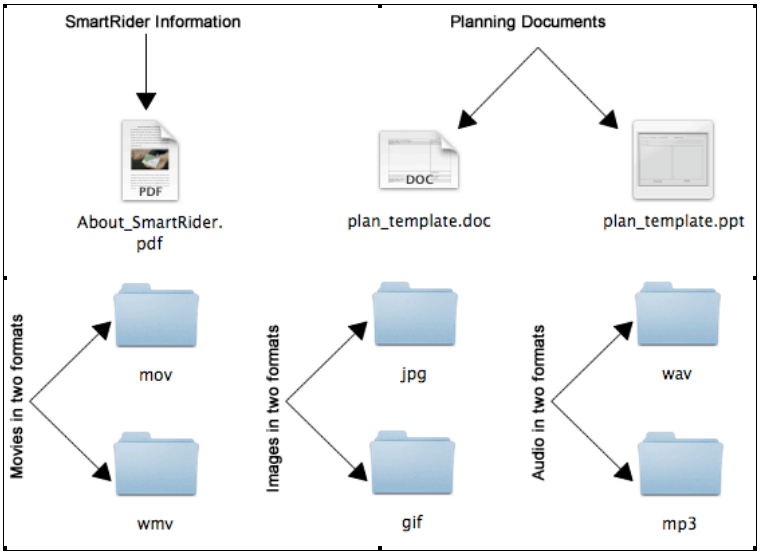

Figure 1: Schema provided on the front page of the AIT exam in the third year explaining the digital resources available on the USB flash drive.