DIGITAL FORMS OF PERFORMANCE ASSESSMENT

C. Paul Newhouse, PhD
Centre for Schooling and Learning Technologies, Edith Cowan University

Abstract

In the developed world a very small proportion of work tasks are done using paper and pen and yet most high-stakes assessment in schools continue to use this primitive technology. While employers and community leaders call for schools to produce students with 21st Century skills and deep conceptual understanding of content the main driver of curriculum and pedagogy, assessment, continues to focus on 19th Century skills and shallow recall of content. In the past it has been considered too difficult to reliably and manageably assess large cohorts using approaches more valid than on paper. The range of maturing digital technologies for handling multimedia now provides opportunities to address this disparity. This session will report on the first year of a three-year project investigating the use of digital technologies to represent student work for high-stakes summative assessment. The project has used a database portfolio system linked to online marking tools and has trialled the use of the comparative pairs method of marking in four senior secondary courses. In 2008 the project involved 21 teachers and their senior secondary classes studying one of Applied Information Technology, Engineering Studies, Italian, or Physical Education Studies.

Introduction

Over the past few years the structure of the curriculum for senior secondary schooling in Western Australia (W.A.) has been changing dramatically. This has led to a large number of new courses being available to students aiming to enter higher education. Some, such as *Applied Information Technology* (AIT), are based on subjects previously designated as non-tertiary entrance while others, such as *Engineering Studies*, are new areas of study. Most of the new courses have major practical components and there is an expectation from students and community that the assessment of student performance will reflect the nature of this learning. For most courses in W.A. students are externally assessed using traditional methods employing predominantly paper and pen technologies. In most cases it is clear that performance on practical tasks cannot be assessed adequately using paper and pen, even in many of the traditional tertiary entrance courses. Many educational researchers (Lane, 2004) argue that traditional assessment only measures knowledge of basic facts and procedures but fails to assess learning processes and higher-order thinking (decision-making, reflection, reasoning and problem solving). Clearly the latter are far more important and useful both to the individual and society in general. The only advantages that paper-based exams have are low cost and ease of authentication.

In the developed world a very small proportion of work tasks are done using paper and pen and yet most high-stakes assessment in schools continue to use this primitive technology. Lin and Dwyer (2006) argue that to date computer technology has really only been used substantially in assessment to automate routine procedures such as for multiple-choice tests and collating marks. They suggest that the focus should be on capturing "more complex performances" (p. 29) that assess a learner's higher-order skills (decision-making, reflection, reasoning and problem solving) and cite examples such as the use of simulations but suggest that this is seldom done due to "technical complexity and logistical problems" (p.28). Thus there is a critical need for research into the use of digital forms of representation of student performance on complex tasks for the purposes of summative assessment that are feasible within the constraints of school contexts.

What is assessed is critical because students tend to focus on, and be motivated by these sections of the curriculum, and teachers tend to ‘teach to the test’. Further, educators are accountable to society

for the outcomes of the use of resources in education, and our society was increasingly expecting that students should demonstrate practical performance not just theoretical understanding. Finally, students are more likely to experience deep learning through complex performance. Therefore as McGaw (2006) explains, this places a responsibility on education authorities to consider strategies to increase the assessment of performance on practical tasks.

> *"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" (McGaw, 2006 p.3).*

Performance-based assessment is not new. Oral and laboratory examinations have been used in European schools and Universities for over a century. In many industries performance-based assessment approaches are used (e.g. pilots). In many high-stakes courses in developed countries performance is, and has been, assessed using observation, interview, portfolio or recording (e.g. USA, UK, Denmark). For example, a recent review of assessment methods in medical education (Norcini & McKinley, 2007) outlines performance-based assessment of clinical, communications and professional skills using observations, recordings and computer-based simulations. In W.A. there has been a history of performance-based assessment in some courses in the Arts. However, the use of performance-based assessment in high-stakes courses has been limited by the costs involved in collecting the evidence of performance and difficulties in ensuring reliable and valid results. Recent advances in psychometric methods and improvements in digital technologies provide tools to assess a variety of performance relatively cost-effectively (McGaw, 2006).

A ground-breaking study aimed at assessing performance, titled e-scape, is being conducted by the Technology Education Research Unit (TERU) at Goldsmiths College, University of London (Kimbell, Wheeler, Miller, & Pollitt, 2007), built upon many years of work on improving assessment in the design and technology curriculum (Kimbell, 2004). E-scape combines three innovations in the assessment of practical performance by representing student work entirely in digital form, collating this work using an online repository and marking it using a comparative pairs judgements technique.

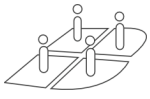
The Project

This paper reports on some of the results of the first year of a three-year project conducted by the Centre for Schooling and Learning Technologies (CSaLT) at Edith Cowan University (ECU) in collaboration with the W.A. Curriculum Council and funded by an Australian Council of Research Linkage grant. The research addresses a critical educational problem, aims for high impact and is building a long-term research agenda. The focus of the study is on the use of digital technologies to 'capture' performance on practical tasks for the purpose of high stakes (i.e. used to determine future opportunities) summative assessment. This may include representations of student 'outputs' or performance processes. The purpose is to increase the authenticity of the assessment.

The study commenced in January 2008 and involved four senior secondary courses of study, *Applied Information Technology* (AIT), *Engineering Studies*, *Italian*, and *Physical Education Studies* (PES). In the first year the study included 21 class-based case studies each involving a teacher and a class in one of the four courses. The number of students involved in each case study ranged from 4 to 26 with 264 students involved overall. This was the *Proof of Concept* stage to be followed in 2009 by a *Prototype*: stage aimed at implementing a prototype assessment task for each of the four courses under 'normal' or typical conditions. Finally, in 2010 the project aims for a *Scalable Product* with the aim to implement each assessment task in a representative sample of schools.

Method

The study is evaluative in nature set within an ethnographic framework in that activity is considered to occur within learning environments where the characteristics of teachers and students and the culture created are critical to an understanding of all aspects of the curriculum and pedagogy, including assessment. The research design required an analysis of both qualitative and quantitative data using



interpretive techniques to investigate the perspectives of the key groups of participants (teachers, assessors, students) with data collected from each group. These data were compiled into case studies in which each case was defined by one class for one course.

The first research task was to develop the common assessment tasks with standardised parameters such as time constraints, level of supervision, level of scripting for implementation, minimum requirements for equipment/resource availability, and types of data to collect. A situation analysis was conducted that considered the characteristics of students, the requirements of the course, the nature of the performance to be assessed, the technologies that could be used to capture this performance, and the characteristics of typical teachers, particularly in relation to the use of these technologies. It was important that the assessment tasks constituted good professional practice, meet the requirements of the course and are reasonably able to be implemented by a 'good' teacher in a real school. However, the aim was to move towards the 'cutting edge' of what is possible. Based on this analysis the structure of each assessment task was formed and examples sought to 'flesh' this out. Teachers were recruited to be involved in implementing the assessment tasks with their classes.

It was decided that for the purposes of this evaluation feasibility would be interpreted using a feasibility framework of four dimensions developed for the British e-scape research (Kimbell, Wheeler, Miller, & Pollitt, 2007). This framework is described in Table 1.

Table 1
Descriptions of the dimensions of feasibility.

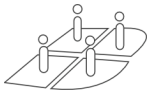
Dimension	Description
Manageability	Concerning making a digital form of assessment do-able in typical classrooms with the normal range of students.
Technical	Concerning the extent to which existing technologies can be adapted for assessment purposes within course requirements.
Functional	Concerning reliability and validity, and the comparability of data with other forms of assessment.
Pedagogic	Concerning the extent to which the use of a digital assessment forms can support and enrich the learning experience of students.

To determine the feasibility of each assessment task a range of types of quantitative and qualitative data was collected including observation in class, a survey of students, interviews with the teacher and a group of students, student work output from the assessment task, and the assessment records of the teacher.

Output from student work on the assessment tasks was collected in digital form and placed in the online digital repository to be available to the markers. The students' work was assessed by two external markers using a traditional analytical 'rubric' method and by a panel of five markers using a comparative pairs method, using sets of criteria developed for each assessment task. In addition the classroom teacher marked the work using whatever approach they wished. For each method of marking a set of marks and rankings were generated both for the entire group of students completing the task and for each separate case study. These were compared using tests of correlation.

The Assessment Tasks

For AIT a hybrid assessment task structure was developed in order to compare the operation of a portfolio with a performance exam. A multi-part reflective process portfolio was developed that included the development of a digital product, the collation of a process document associated with the digital product, and the presentation of two other digital artefacts. A three-hour performance tasks exam was developed that included a set of questions reflecting on the development of the digital product for the portfolio. The majority of the performance tasks exam involved students designing, producing and evaluating a logo and brochure associated with a design brief challenge. The tasks



were completed using school computer workstations in laboratories and USB flash drives.

For Engineering Studies the assessment task was a series of timed specified activities, which took students from a design brief to the construction of a model over a period of three hours. Each student had an ASUS eeePC computer, a web-cam and a USB flash drive on which they compiled their portfolio within a pre-formed Filemaker database template. Input consisted of text, graphics through a camera, voice and video through a web-cam. A researcher or the teacher coordinated the activity by showing a presentation projected onto a screen. This was used for managing the time students had on each activity. The task involved the design and modelling of a solar water heater for a rural developing country context.

For Italian the assessment task comprised a folio and an oral presentation. The folio was a series of tasks and activities to show development of ideas and preparation for the presentation. It included a map activity, retrieval chart and question answers, brainstorm, fact sheet, word-processed reflection on practice talk and one minute voice recording. The oral presentation required the student to prepare and deliver a two-minute talk focussed on a local area/WA destination, providing information such as, features of the area, recreational activities and cultural events. The oral presentation was recorded using a digital video camera and remote lapel microphone.

For PES the assessment task was a performance tasks exam including response questions. The task was completed in four sessions. In the first session students were set a tactical challenge situation for their sport and asked to analyse the challenge and propose solutions by typing responses to questions using a pre-formed Filemaker database template. In the second session students undertook four different drills to demonstrate skills that were relevant to the challenge. In the third session each students were in modified 'game' situations to demonstrate their solutions to the tactical challenge. Performances in the second and third sessions were video recorded using a remote control multi-PTZ-camera system. In the final session students viewed videos of their performances and responded to reflective and evaluative questions. This was completed using a pre-formed Filemaker database template with online links to digital video files stored on a remote server.

Data Analysis

A range of types of data was collected. These data were initially analysed separately for each case study and then combined for each course.

Observations, Interviews and Surveys

All classes were observed completing each of the components of the assessment tasks. On completion of the tasks the teacher was interviewed as well as a representative group of students from the class. In general it was found that the AIT and Engineering teachers (except for one) and students were positive about the assessment task and its implementation and perceived few difficulties. The PES teachers and three of the groups of students were also positive. The Italian teachers were divided in their views and the students tended to be negative towards the recording of the oral presentation.

Students completed a questionnaire consisting of 57 closed response items and two open-response items. A number of scales were derived from combining items from the questionnaire. Descriptions of the scales and means for the four courses are shown in Table 2 below. In general students in AIT and Engineering found completion of the assessment tasks relatively easy, this was less so for PES and Italian. As would be expected the mean ICT skills score was lower for these two groups of students as was their attitude towards using computers and the amount of time they used computers at school. As would be expected the AIT students indicated using computers considerably more than the other three groups of students.

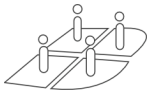


Table 2
Descriptions and mean statistics for the scales based on items from the student questionnaire.

	AIT	Eng	Ital	PES	Description
eAssessE	3.2 (0.4)	3.2	2.7	2.9	Ease of completion of the exam. Score between 1 and 4
eAssessP	3.2 (0.4)				Ease of completion of the portfolio. Score between 1 and 4.
Apply	2.4 (0.4)	2.2	2.4	1.9	Application of computer use. Score between 1 and 3.
Attitude	2.6 (0.3)	2.6	2.4	2.4	Attitude towards using computers. Score between 1 and 3.
Confidence	2.7 (0.4)	2.8	2.5	2.6	Confidence in using computers. Score between 1 and 3.
Skills	3.3 (0.5)	3.1	2.9	2.9	Self-assessment of ICT skills. Score between 1 and 4.
SCUse	96 (62)	49	36	41	Estimate of time in mins/day using computers at school.

Note: There was little difference between the standard deviations for the four courses and therefore they are not quoted here.

Marking and Methods of Marking

The analytical method of marking using rubrics was successfully implemented for the assessment tasks in all four courses. The comparative pairs method of marking was successfully implemented in all four courses but in Italian this was only done for the oral presentation and for AIT only for the performance tasks exam. For the comparative pairs method a set of three criteria was developed for each assessment task as well as a holistic criterion to support assessors in making four judgements for each pair. For both methods of marking the process was supported by the use of online digital marking tools and an online portfolio system created using directory and file names on a remote server accessed through a browser using Filemaker online.

Markers judgements were initially captured in Filemaker data files and then exported. The correlation coefficients for all the courses are reported in Table 3. The marks from the two assessors using the analytical method of marking and resulting rankings were highly correlated in all the courses except for Engineering. There was at best only moderate correlation with the teacher marks but strong correlation between the two methods of marking for all courses. The lack of correlation with teacher marks was likely to be due to a combination of teachers using different criteria, most classes being small samples and teachers taking into account background knowledge of the students or tasks. The Separation Index (SI) values from the Rasch analysis of the comparative pairs marking were all high (between 0.92 and 0.95) indicating a high reliability of the scores generated.

Table 3
Correlations between assessors, assessors and teachers, and methods of marking.

			Correlations Between Markers or Methods of Marking					
			Analytic Assessors		Analytical and Teacher		Analytic and 'Pairs'	
Course	N	SI	Marks	Rankings	Marks	Rankings	Marks	Rankings
AIT	115	0.93	0.89**	0.91**	0.32	0.30	0.73**	0.71**
Engineering	68	0.92	0.43**	0.42**	0.54**	0.66**	0.78**	0.72**
Italian	35	0.95	0.93**	0.87**	0.20	0.25	0.70**	0.63**
PES	39	0.92	0.87**	0.87**	0.52**	0.57**	0.89**	0.88**

** p<0.01 (2-tailed) * p<0.05 (2-tailed). SI = Separation Index for comparative pairs marking

Results

The results for each of the four courses are provided separately.

Applied Information Technology

Both the portfolio and exam were implemented for all seven classes, however, the extent to which

students completed all components of the portfolio varied considerably. The exam was completed by all students with almost no technical difficulties evident apart from the recording of sound for three classes. Almost all students indicated a preference for the two forms of assessment and that both provided a good assessment of practical performance. They commented on the ease of working on the computer compared to working on paper … correcting errors, speed of writing, amount of writing, speed of action and physical comfort. When pressed, their major concern was malfunctions of systems during exam.

The manner in which the portfolio was implemented varied somewhat between classes mainly on the extent to which this was included as a part of the school-based assessment. Generally the product requirements provided adequate scope for students to demonstrate their capability. The process document varied considerably in quality with some lack of understanding of technology process. The reflective question component of the exam did highlight some discrepancies on portfolio products and process document for some students. Typing into the word-processed document was efficient but one-hour was too long and the results were of limited value. The performance tasks component of the exam provided scope for demonstration of capability. All students completed most requirements without difficulty, apart from sound recording, and the inclusion of graphs was generally poor. There were significant moderate correlations between the scores on the portfolio and exam (around $r = 0.5$).

Engineering Studies

All five classes of students successfully completed the assessment task that was implemented in almost identical fashion in each case. Almost all students preferred the assessment of their engineering performance through this means rather than a paper examination. They found the experience engaging and enjoyable, and felt that it more accurately reflected the nature of the engineering course they were studying. Many students seemed to have a natural affinity with the range of technology used in the examination. Their two main concerns were with the size of the keyboard on the computer resulting in difficulty in typing, and the low resolution of the web-cam resulting in poor representation of their sketches. The difficulty in interpreting the development taking place in the series of sketches because of this poor resolution was also highlighted by a number of the assessors. Although there were not always significant correlations between the markers scores using the analytical rubric based approach, overall the correlation was low but significant at 0.43.

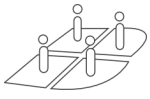
Italian

Each of the four classes completed the oral presentation component of the assessment but varied in the extent to which the portfolio was completed. The teacher facilitated the completion of the portfolio and assisted a researcher in facilitating the video recording of the oral presentations. The portfolio tasks were only fully completed in one class. In all four schools students were recorded (video and audio) completing the oral presentations. However, there were some variations including talking for 5 minutes instead of 2 minutes and the use of presentation software (Powerpoint) during the presentations. There were inconsistencies in the way that an unseen question was delivered.

The students indicated little experience in doing assessments using digital technologies. About two thirds of the students agreed that, overall, digital technologies were good tools for parts of the Italian exam. About half agreed and half disagreed with the assertion that “I was able to show what I can do in the exam”. Seventy percent of the students disagreed that it was better doing the oral exam with digital technologies than face-to-face with an examiner. However, the students appreciated being able to analyse and critique their own performances through reviewing the video recording of their performance. They almost unanimously reported that they felt very nervous being video recorded and therefore felt that they were not able to perform at their best.

Physical Education Studies

Each of the four classes involved a different sporting context: rugby union, soccer, swimming and volleyball. All four sessions of the performance tasks exam were completed successfully with all four classes. For two classes the video sessions had to be repeated to included absent students and for three classes there were difficulties in accessing laboratories for the computer-based components.



Generally, the task was regarded by students as an appropriate means of assessment for the course. The task effectively encompassed conceptual, practical and reflective aspects and was able to be adapted for application and implementation in varied sporting contexts. Students identified that a sole reliance on text as a means of response to questions was limiting and recognised that it would be advantageous to also be able to use graphics/drawing tools. The practical components of the task were implemented effectively. These were designed to enable all students to have the opportunity to demonstrate their performance abilities. The task has the scope to be undertaken over a more compacted time frame and this would be beneficial from a student perspective.

Conclusion

To address the research questions, the results of the data analysis were interpreted within the four dimensions of the *Feasibility Framework* as described in Table 2.

Manageability

It was possible to implement each of the assessment tasks successfully with a class. There were few, if any, logistical difficulties for production or performance tasks exams and only time constraints and difficulties in managing school-based assessment requirements that limited the effectiveness of portfolios. There was little evidence that any students were adversely affected by a lack of ICT experience or capability. There were some difficulties in markers accessing student work from within particular network firewalls or using slow Internet access.

Technical

Overall there were no significant technical difficulties that could not be relatively easily overcome. Student work was collected either on a class DVD, USB drives or cameras and uploaded to a University server. There were no difficulties with these processes. In the AIT, Italian and Engineering Studies courses some degree of maintenance was required to prepare student work for markers. For PES researchers controlled all digital capture (video and FileMaker database on USB drives). The most was required for the AIT portfolio as students used a range of software and digital formats. For PES some video capture needed enhancing, particularly with the need for an underwater shot for swimming, a close-up shot and clearer shots on overcast days.

Functional

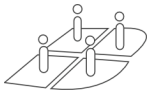
The students perceived the assessment tasks to be authentic and meaningful and, apart from in Italian, preferred the task to a written exam. In almost all cases the teacher perceived the assessment to be more authentic than a paper-based exam. For all four courses the assessment task was structured permitting a good range of levels of achievement to be demonstrated. In the AIT exam there may have been limitations to the opportunity to demonstrate higher-level achievement due to the nature of the tasks involved. External assessors used analytical marking rubrics that resulted in highly correlated marks and rankings except for in Engineering where they were only moderately correlated. The comparative pairs method of marking was successfully implemented with resulting highly reliable scores.

Pedagogic

In almost all cases for AIT, Engineering and PES the assessment matched general pedagogy for the course and was viewed positively by teachers. This was not the case for one AIT teacher and one Engineering teacher. In all four courses the quality of work was highly dependent on the class the student was in probably reflecting differences in capability of the students and pedagogical approaches by the teachers involved. In Engineering, in particular, students tended to feel that they need more time and more flexibility in what they did and the order in which components were completed.

Constraints and Conclusions

The first year of the study identified relatively few constraints to the use of the digital forms of



assessment implemented in the sample of schools. The main constraint was the extent to which the teacher was able to include the task within their school-based assessment framework. In PES school timetabling with short time periods was a constraint for three cases. Student lack of familiarity with the Italian portfolio and AIT process document were constraints. In PES and AIT there were some ICT infrastructure constraints in a few schools but in general these were relatively minor.

Overall the benefits outweighed the constraints. Analysis of all the data has allowed refinements to be made to the assessment tasks with a view to the potential for their implementation with many students in Western Australia. Lessons learned should then guide implementation of similar assessment tasks for a wider range of senior secondary courses.

References

Kimbell, R. (2004). Design & Technology. In J. White (Ed.), *Rethinking the School Curriculum: Values, Aims and Purposes.* (pp. 40-59). New York & London.: Routledge Falmer.

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.

McGaw, B. (2006). *Assessment to fit for purpose.* Paper presented at the 32nd Annual Conference of the International Association for Educational Assessment., Singapore.

Norcini, J. J., & McKinley, D. W. (2007). Assessment methods in medical education. *Teaching and Teacher Education, 23*(1), 239-250.

Acknowledgement

The theory discussed in this paper and the research upon which it is based are as a result of the work of a research team organised by the Centre for Schooling and Learning Technologies at Edith Cowan University. The team was led by Paul Newhouse and John Williams and includes researchers Dawn Penney (University of Tasmania), 'Chirp' Lim, Jeremy Pagram, Andrew Jones, Mark Hackling, Ron Oliver, Russell Waugh, Martin Cooper and Alistair Campbell, and a number of research assistants.