Authentic Assessment in Performance-based Subjects

Assoc Prof P John Williams

Prof Dawn Penney

Faculty of Education,

The University of Waikato

Abstract

This paper reports on a three-year study conducted in Western Australia, which commenced in January 2008, and was completed by December 2010. It concerns the potential to use digital technologies to represent the output from assessment tasks in two senior secondary courses: Engineering Studies and Physical Education Studies.

The general aim of this study was to explore the potential of various digitally based forms for external assessment for senior secondary courses in terms of manageability, cost, validity and reliability. The problem being addressed in this research is the need to provide students with assessment opportunities that are authentic, where many outcomes do not lend themselves to being assessed using pen and paper over a three-hour period, and that are also able to be reliably and manageably assessed by external examiners. These two courses both have a significant performance-based component, and a certain level of dissonance results when students have performance expectations, and teachers teach to the theory examination in an attempt to ensure high pass rates.

In Engineering Studies, a computer-managed examination was designed that consisted of a design task that was broken down into a number of timed activities. Students were paced through each activity, recording their input in the form of a portfolio. Input consisted of text, graphics through a camera and voice, and the exam outputs were uploaded to an online repository for marking by external assessors.

In Physical Education Studies, a digitally based examination was designed that incorporated four inter-related components. Two of these required computer-based responses and two were video-recorded practical performance components. Digital output from all parts of the task was collated into evidence portfolios and uploaded to an online repository for marking by external assessors.

This paper focuses specifically on issues of authenticity in relation to the examination tasks designed for each course. It discusses longstanding concerns relating to authenticity associated with courses with a strong performance dimension in which there tends to be a dissonance between the teaching and learning of theory and practice. Ways in which digital technologies were utilised to address these issues are critically examined in this research.

Introduction

The context for the study reported in this paper is the senior secondary curriculum in Western Australia (WA), a review of the curriculum in 2002 (Curriculum Council of WA, 2002), resulted in the development of some 50 new senior secondary courses in a range of subject areas, with all courses designed to service multiple career pathways. Notably, students could study and use their assessment results from any of the new courses for university entrance scores. For some areas of the curriculum, including Physical Education and Engineering, this was a new development, as they had not previously had the status of being a "university entrance subject" in WA.

In addition, some of these new subjects were performance based, for example, Engineering, which is based on design performance, and Physical Education, which is based on physical performance. This introduced a number of novel dimensions to the learning and assessment process. Primarily, there is an expectation by many stakeholders that the assessment processes are rigorous, authentic and reliable. This was immediately problematic for performance-based subjects as the prevailing assessment regime was dominated by three-hour summative examinations. Many educational researchers argue that these forms of traditional assessment fail to assess processes and higher order thinking skills (Lane, 2004; Lin & Dwyer, 2006), and so the authenticity of the form of assessment (written examination) to measure a student's performance was questionable. In the case of Physical Education, Hay (2006) has argued that "authentic assessment in PE should be based in movement and capture the cognitive and psychomotor processes involved in the competent performance of physical activities" (p. 317) and furthermore, do so via assessment approaches and tasks that acknowledge and reflect the inter-relatedness of these processes in physical performance.

This research also acknowledged that teachers tend to teach to the summative assessment (Barnes, Clarke, & Stephens, 2000; Lane, 2004; Ridgway, McCusker, & Pead, 2006). If external examination took a traditional written form, the danger was that teachers would tend to ignore the basic performance nature of the subject they were teaching, and design learning experiences and utilise pedagogies primarily directed towards the development of more effective examination writing skills. This is quite explicable in a context where teachers are expected to teach their students in such a way as to maximise their chances of success, and in which performance outcomes of schools are becoming an increasingly significant and more public measure of a school. McGaw (2006) discussed this, noting the "risk that excessive attention will be given to those aspects of the curriculum that are assessed" and that consequently "risk-taking is likely to be suppressed" (p. 2). He goes as far as to argue that "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). The potential outcome of this dissonance between the nature of the subject (performance based) and the

form of assessment (written examination) is therefore inauthentic assessment and inappropriate pedagogies. This provided a strong rationale for exploring alternative methods of assessment in WA.

A corollary of this dissonance is a potential separation of theory and practice in the presentation of the subject to the students. In senior secondary schooling, the inclusion of a distinct practical component to an examination with little or no clear links to theoretical aspects of a course can clearly accentuate a dichotomous view. Equally, the absence of a practical or applied component to an examination can reaffirm such a view, signalling the higher status of theoretical knowledge. A further challenge in WA was, therefore, to develop ways of assessing performance-based subjects in a manner that promoted an applied and integrated perspective, and that did not present theoretical and practical elements as disconnected. As we discuss further below, this also related to concerns for authenticity in assessment and specifically, external examinations.

Concurrent with the above curriculum developments in WA were advances in psychometric methods, and improvements in digital technologies internationally. From the 1990s, significant developments in computer technology have seen the emergence of low-cost, high-powered portable computers, and improvements in the capabilities and operation of computer networks (e.g. intranets and the accessibility of the Internet). These technologies have appeared in schools at an escalating rate and provided the mechanism for exploring alternative methods of assessment in senior secondary schooling in WA.

While there are numerous examples of the use of digital technologies in assessment across a range of subject areas, their use in high-stakes school-level performance assessment is relatively rare, for a range of reasons that tend to relate to feasibility concerns. Initially, concerns about cost, logistics and technical reliability were foremost (Lin & Dwyer, 2006), but Dede (2003) suggests that the barriers to using digital technologies to support alternative forms of assessment are not so much technical or economic as "psychological, organizational, political and cultural" (p. 9). That is, participants, educators, leaders and community members are not adequately convinced of the efficacy of computer-supported or -based assessment. The familiar mode and format (written, paper and pen, supervised mass examinations held on a designated day and time) are largely accepted as providing a legitimate and reliable means of assessment, with alternatives and specifically, digital assessments, viewed with hesitancy and some scepticism. To some extent this is due to a lack of understanding or knowledge, but largely, it indicates the need for compelling research findings.

Conceptual Framework

In order to investigate the use of digital representations to deliver authentic and reliable assessments of performance, this study brought together three key innovations:

1. The representation in digital files of the performance of students doing practical work.
2. The storage of digital representations of student performance in an online repository so that they are easily accessible to markers.
3. Assessing the digital representations of student performance using both standards-referenced judgement and the paired comparison judgement methods with holistic and component criteria-based judgements.

Below we explain each innovation further as it was applied in the research project. While each of these innovations is not new in themselves (Messick, 1994), their combination applied at the secondary level of education was new. Apart from Kimbell and colleagues, (2007) work at the University of London, focusing on examination assessment in design and technology, science and geography at GCSE[1] level, there was no known precedent.

There are, however, many types of digital portfolios used internationally for a range of purposes. In a review of e-assessment, the digital portfolio is recommended as a "way forward" in the high-stakes assessment of "practical" work in that Information and Communications Technology (ICT) "provides an opportunity to introduce manageable, high quality coursework as part of the summative assessment process" (Ridgway, et al., 2006).

[1] General Certificate of Secondary Education is the examination taken by students at the end of compulsory secondary schooling in England (aged 15–16 years).

There have also been trials on a variety of computer-based exams, particularly over the past decade, as computer systems have become more robust, networks more reliable, and software more flexible. For example, in the Canadian provinces of Alberta, British Columbia and Ontario on-screen online exams have been used for high-stakes assessment for several years across a considerable range of subject disciplines (Carbol, 2007). In Norway students use government-provided notebook computers to complete examinations across a range of disciplines (BBC, 2009). In the UK, in the e-Scape project, students use handheld or notebook computers to respond to questions and capture audiovisual evidence of activity in design and technology, science and geography (Kimbell, Wheeler, Miller, & Pollitt, 2007). Wiegers (2010) reports that in The Netherlands computer-based exams have been used to some extent since 2000 with a prediction that in 2011 37% of national exams will be computer-based. Some of these are "non-linear, interactive, open-ended assessment tasks leaning heavily on use of multimedia" (p. 2) and are used to examine skills not able to be assessed on paper, to increase alignment with life requirements, increase flexibility in delivery of assessment and to reduce workload. Although Wiegers does not discuss exams involving digital creation of artefacts, he does suggest pre-conditions and an implementation procedure for computer-based exams. This procedure involves six years to full system-wide implementation with the first four years culminating in a full-scale pilot (the endpoint reached by the present study).

The WA study reported here investigated the use of digital forms of representation of student practical performance for summative assessment and specifically, external examination in senior secondary schooling. Digital representations of student performance in tasks that comprised multiple parts were combined within an online repository. An evidence portfolio, in a digital format, was thus created for each student. A key feature was that the portfolios could be accessed from anywhere, enabling markers from different jurisdictions to be involved, enhancing consistency of standards.

As indicated above, the project utilised two methods of marking. Both were conducted fully online, using the same digital portfolios of student evidence. For the standards-based analytic assessment, rubrics were developed and markers were required to make a series of judgements about performance in relation to the criteria and standards specified.

The paired comparison judgement method of marking was implemented using holistic judgements. As the name suggests, this method requires a direct comparison of two students' performance in the task, with each student's full evidence record from all parts of the task available to assessors. Pairings are automatically generated and assessors are required to decide which of the two students being compared has performed better in relation to the set criteria. While Pollitt (2004) describes the method as "intrinsically more valid" and better than the traditional system, he believes that without some ICT support it has not been feasible to apply due to time and cost constraints. Pollitt (2004) has therefore suggested that further research is required to determine the appropriateness of the method and whether "sufficient precision can be achieved without excessive cost" (p. 16). McGaw (2006) believes that such methods being supported by digital technologies should be applied in public examinations.

Method

The general aim of this study was to explore the potential of various digitally based forms for external assessment for senior secondary courses in WA. The problem being addressed was the need to provide students with assessment opportunities that were authentic, where many outcomes do not lend themselves to being assessed using pen and paper over a three-hour period, and that were also able to be reliably and manageably assessed by external examiners. That is, the external assessment for a course needs to accurately and reliably assess the course outcomes without a huge increase in the cost of assessment. In WA, a factor to consider in relation to cost of examinations is the vast geographical spread of schools and students. The assessment tasks and technologies needed to be able to be implemented across the state in diverse school contexts.

The project addressed four of the new senior secondary school courses that each encompassed a performance dimension: Applied Information Technology, Engineering Studies, Physical Education Studies, and Italian. For each course, the research involved a three-year, phased development. The first year of the study was a "proof of concept" project to explore the feasibility of particular digitally based formats for external assessment appropriate to each course. Feasibility was investigated within a framework consisting of the four dimensions: technological, pedagogic, manageability, and functionality. The second year of the study focussed on developing a full and robust prototype of a digitally based assessment for each course. In the third year the prototype was scaled up to be implemented in a larger sample of representative schools.

During each year of the project a range of types of quantitative and qualitative data were collected including observation in class, a survey of students, a survey of the teachers, interviews with the teachers and a group of students, student work output from the assessment task, and the assessment records of the teacher. Quantitative and qualitative data were also generated from the standards-based and comparative pairs marking. Markers were able to make written comments during the marking process about aspects of the marking interface and/or specific judgements and interviews were also conducted with markers.

The following sections report on the research undertaken in relation to two of the four courses: Engineering Studies and Physical Education Studies.

Engineering

The Engineering Studies course consists of a core which all students address, covering the three areas of engineering design and process; common engineering principles, structures and systems; and enterprise, environment and community. Students can then specialise in one of three options: materials, structures and mechanical systems or electrical/electronic or systems and control.

> *Engineering Studies provides a focus on design through exciting creative, practical and relevant opportunities for students to investigate, research and present information, design and make products and undertake project development. These activities provide students with opportunities to apply engineering processes, understand underpinning scientific and mathematical principles, develop engineering technology skills and to understand the interrelationships between engineering projects and society. (Curriculum Council of WA, 2007, p. 3)*

An Extended Production Exam is the completion, under "exam conditions", of one practical assessment task that incorporated a full set of processes (e.g. design process, scientific investigation) and centred on one major scenario. Examples were found locally, nationally and internationally of performance on practical tasks being assessed through an extended production, or small project, under exam conditions. However, most did not involve the use of digital technologies. The most comprehensive example was that of Kimbell et al. (2007) in the UK where students spent two consecutive mornings of three hours duration each working on a structured design activity for the production of a pill dispenser. All student work output was collected digitally using a networked Personal Digital Assistant (PDA) device and local server. An adaptation of this form was used for the engineering course. In order to be appropriate for all students, the production examination content was designed with the school teachers, and had to focus on the core of the Engineering course, so it would be relevant to all students regardless of their specialisation. The third year of the study involved eight teachers and their eight upper secondary (Year 11) classes containing 94 students, from city, country and remote localities.

A template was used to present a set of tasks to the students and so scaffold responses to a design brief. The brief was to design a product that would enable someone stranded on a beach with no drinking water to use the power of the sun to produce drinkable water from sea water using a limited range of available materials. The tasks involved the students in developing a number of design iterations in response to a range of stimulus material. Throughout the examination they produced four design sketches, each in response to new stimulus material, such as a materials data table. The expectation was that the students would progress their design over the four iterations. They used a webcam to photograph their sketches, and then annotated them online. They also reflected on their design in a video; devised evaluation criteria in a table; discussed a mass production application and evaluated the impacts of large-scale desalination plants. This was input through the types of pages illustrated in Figure 1.

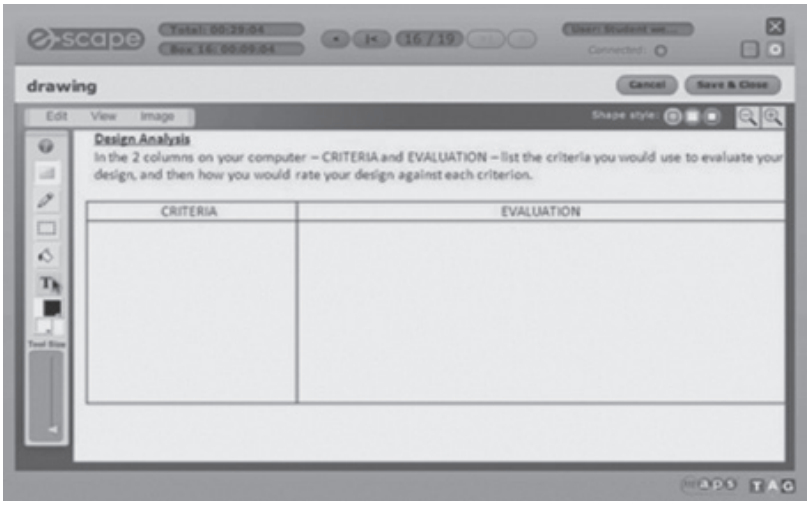

Figure 1. Two samples of pages of the examination as presented to students.

The task was managed in three different ways across the eight schools, but the appearance of each method on the students' computers was the same:

1. Intranet. In this scenario all the students were issued with ASUS EeePC mini computers (Figure 2) which were wirelessly linked with the research facilitator's computer. The facilitator was then able to monitor each logged-in student's progress throughout the examination tasks, and progress all the students on to the next task at the same time.

Figure 2. An ASUS EeePC computer with stand to hold external USB camera

2. Live. In this scenario the students were working on the school computers and logged on live to the examination server. The research facilitator was also logged on as the examination manager and was then able to monitor each logged-in student's progress throughout the examination tasks and move the students on to each task when the time expired.
3. USB. In this scenario the examination was accessed by the students through logging on to a USB drive which was inserted into a school computer. This did not enable centralised control by the research facilitator of the examination administration, but students were able to progress through the exam at their own rate.

The outputs for the 94 students were uploaded to the online repository. The students' work was marked by two external assessors using a standards-referenced rubric. At the same time each teacher marked his own students' work using his own method. There was a moderately significant correlation between the two external assessors with correlation coefficients of 0.53 ($p<0.01$). This would tend to indicate that the scores were reasonably reliable but that the marking guides maybe could have been more explicit. There were also moderately significant correlations between the average score of the external assessors and the scores awarded by the teachers for the examination ($r=0.64$, $p<0.01$). This suggests that, in spite of the differences in assessment criteria, student competence was recognised consistently by both assessors and teachers.

The exam output for the 94 students was also assessed using comparative pairs judging. Ten assessors made 473 comparative judgements between the exam outputs of pairs of students. The pairs to be judged were dynamically generated by the Pairs Engine software[2]. The presentation of the students' portfolio to the assessor, and screen for making judgements is illustrated in Figure 3. The assessors were teachers and members of the research team, two being also involved in the standards-referenced marking.

[2] The Pairs Engine software was developed by TAG Learning in the UK and was designed to present pairs of portfolios to the assessors on the basis of real-time statistical calculations of which pairs needed judging in order to enhance reliability.

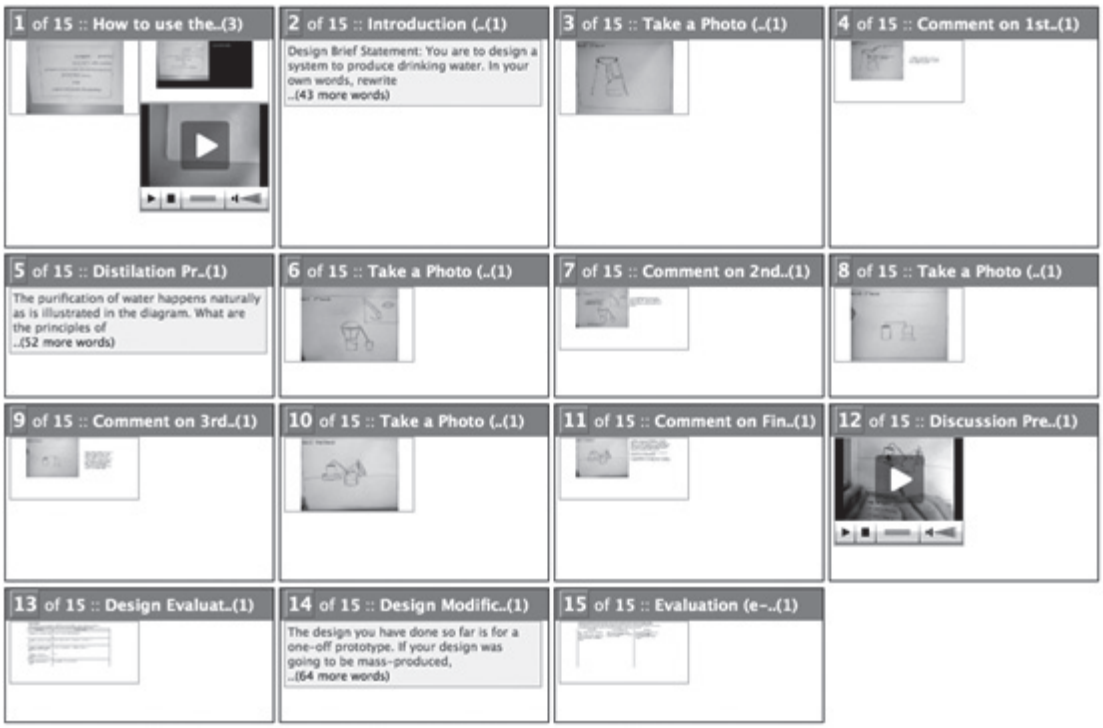

Figure 3. Screen displays from the Pairs Engine used for comparative pairs judging.

An holistic statement was developed to guide assessors in making judgements:

- Was the student able to progress from their initial idea, in response to a range of stimulus and activities, to a satisfactory solution in a manner that clearly communicated the rationale for doing so?

The holistic criteria are made up from the following specific questions:

1. Did the student consider the context (people and environment) in their design?
2. Did the student show idea progression throughout the task?
3. Was the design creative?
4. Was communication clear and relevant—graphics, text and oral?

The reliability coefficient reached 0.9 after 12 rounds of marking. There was a moderate and significant correlation ($r=0.46$, $p<0.01$) between the scores generated by comparative pairs marking and the score determined by analytical marking.

Physical Education

The Physical Education Studies course (Curriculum Council of WA, 2010) encompasses three areas of content: (i) movement, skills, strategies and tactics; (ii) physiological dimensions; and (iii) social dimensions. All units of study (usually two per year) address each of the content areas, with the intention that a multi-disciplinary approach to physical education will be adopted and that students will develop progressively more sophisticated knowledge in, through and about physical education (Arnold, 1979, 1988) and thereby extend achievement in relation to the course outcomes. The Physical Education Studies course outcomes reflect that the course and areas of content seek to address links between theoretical and practical dimensions of knowledge or "ways of knowing" in physical education (see also Brown & Penney, 2011; Thorburn, 2007), and ensure that Physical Education Studies is a "performance-based" course. The four course outcomes are: Skills for physical activity; Self-management and interpersonal skills for physical activity; Knowledge and understanding of movement and conditioning concepts for physical activity; and Knowledge and understanding of sport psychology concepts for physical activity (Curriculum Council of WA, 2010).

The Physical Education Studies course allows for variation in the physical activity contexts that can be utilised as the focus and context of learning in any unit of study. In this respect the course was designed to provide teachers with flexibility to utilise a physical activity context relevant to particular groups of students in schools and local communities that vary in terms of the activities that are meaningful and available.

The task designed for the research project reflected these course characteristics. The Curriculum Council of WA were concerned first and foremost for the research to effectively engage with the "movement, skills, strategies and tactics" course content. In relation to the course outcomes, our focus was therefore Outcome 1: Skills for physical activity; and parts of Outcome 3: Knowledge and understanding of movement and conditioning concepts for physical activity. An integrated task, comprising four parts, was designed to reflect these foci and to accommodate varied physical activity contexts. The task centred on a tactical problem, such as a corner kick in soccer, or sideline pass in netball, or a specific race scenario in swimming. Each part of the task generated digital evidence.

- Part 1. This was computer-based, undertaken in a computer laboratory or using lap-top computers in a classroom. Students were required to respond to a series of structured questions relating to a tactical problem in a specific activity context. They could use text and drawing tools in their response, with graphics of court/field situations available to annotate. Text and graphic responses thus formed the first component of the evidence portfolio.
- Part 2. This part was field-based and was video recorded using a multi-camera set up to capture performance from different angles. Cameras purchased from Vizcom were used and are illustrated in Figure 4.
- Students were required to perform four specified skills pertinent to the tactical problem in controlled conditions. For example, in netball, chest pass, shoulder pass, shadowing, and moving away from an opponent to receive a pass. Students rotated through the skill-drills and video recordings of their performance provided the second component of the evidence portfolio. The video feed from three cameras was automatically synchronised at the time of capture to produce a single multiple-view split-screen recording of each student's performance. Figure 5 below shows examples of the camera set-up for different activities. Cameras were linked to a lap-top computer and could all be remotely controlled from the lap-top by one of the researchers.

Figure 4. PTZ Camera used with remote-controlled multi-camera system

- Part 3. This part was also field-based and video recorded. Students participated in modified game/competition situations designed to simulate the tactical problem and require them to apply skills, knowledge and understanding in a dynamic performance context. Video recordings of student performance provided the third component of the evidence portfolio.

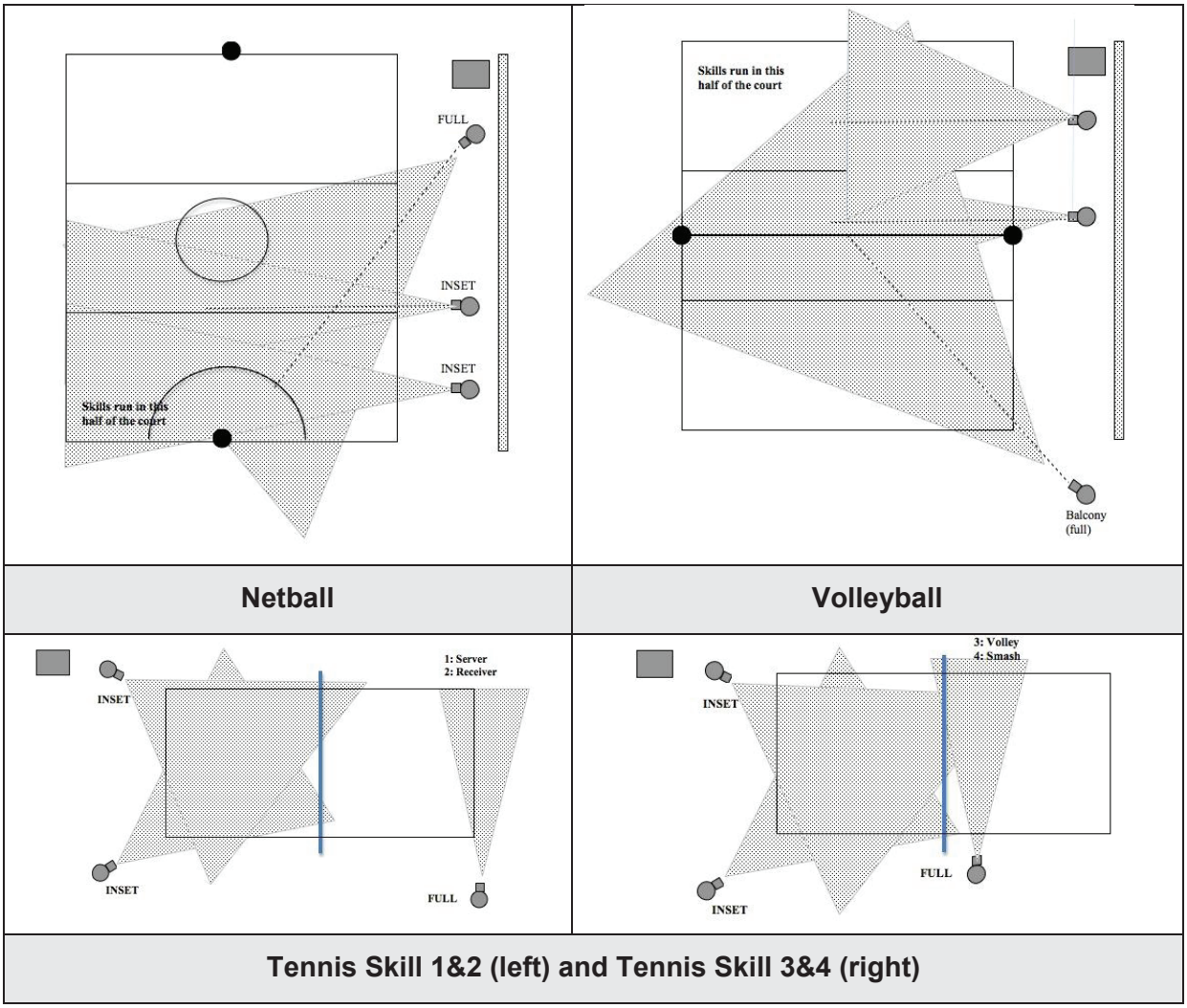

Figure 5. Camera positions for the drills for different physical activity contexts.

- Part 4. This was computer-based, undertaken in a computer laboratory or using lap-top computers in a classroom. Students had access to all of their video footage from parts 2 and 3 for this part of the examination. They were required to respond to a series of structured questions challenging them to critically reflect on their performance. Once again, they could utilise text and drawing tools in their responses. Text and graphic responses thus formed the final component of the evidence portfolio.

The digital representations of student work generated from each part were uploaded to an online repository to be accessed by assessors, with all evidence for each student collated into an individual record. All marking was undertaken online and the same digital evidence was used in both standards-based analytic assessment and comparative pairs assessment.

The third year of the study involved 6 schools, 12 teachers and 11 classes of Year 11 and 12 students taking the Physical Education Studies course. Across the schools, teaching and learning utilised netball, volleyball, swimming, cricket, tennis and soccer. The four-part task was therefore adapted for each activity context and in the case of cricket, two variations were developed to suit students specialising in batting or alternatively, bowling.

The exam outputs for 148 students were uploaded to the online repository. The student exam outputs were marked firstly using the standards-based method, with an online marking tool incorporating a marking key based on the assessment criteria developed for the task and to align with the Physical Education Studies syllabus document. Each student's performance was assessed by two external assessors who were either general Physical Education experts or experts in the sporting context.

The class teacher also marked the students' examination work using their own analytic marks-based system and was asked to provide their semester grade for students. For each assessor and for the teacher the students were ranked so that comparisons could be made.

Generally there was a good spread of scores for both assessors for almost all components of the assessment task. The scores awarded by the two assessors were significantly but only weakly correlated ($r=0.01$, $p<0.01$) and similarly, for the resulting ranking of students. Scores resulting from teacher marking were also significantly but weakly correlated.

Comparative pairs judging involved the assessment of the examination output for 108 students. (Some sets of student output were not included because one or more videos were identified in the analytic marking as too difficult to readily view for judgement purposes). Twenty assessors each completed a set of comparisons between the exam outputs of pairs of students using the Pairs Engine. The pairs to be judged were dynamically generated by the software. All judges were general Physical Education experts, four being involved in the standards-based analytical marking. One holistic and three subsidiary assessment criteria were detailed for the comparative pairs process, with one judgement required, encompassing all criteria.

The holistic criterion judgement about performance addressed students' ability to make informed decisions in, and about, performance situations and to effectively execute responses to those situations. Execution of responses is acknowledged as encompassing skill execution and implementation of strategies and tactics. In making a holistic judgement the assessors were instructed to keep the following specific criteria in mind:

1. Knowledge and understanding of strategies and tactics: conceptual understanding of game/performance situations, the problems that specific situations pose and viable solutions that may be enacted in responding to situations.
2. Execution of movement skills: Execution of the skill is deemed to encompass the essential movement that constitutes preparation for and following through from a point deemed that of enactment of a skill.
3. Application of strategies and tactics: the application of strategies and tactics as demonstrated by students in a live performance context.

Assessors could access all evidence from students' evidence portfolios and could make comments about each portfolio for reference. Figure 6 shows screen shots from the Pairs Engine interface.

Figure 6. Screen displays from the Pairs Engine used for comparative pairs judging.

It had been decided to stop marking once the Cronbach Alpha Reliability Coefficient was above 0.95. This occurred after the 13th round of marking, and a total of 710 judgements, when this coefficient was 0.958. Overall the system only identified 46 (6.5%) of the 710 judgements made that appeared to be seriously inconsistent.

There was a moderate to strong and significant correlation ($r=0.73$, $p<0.01$) between the scores generated by comparative pairs marking and the score determined by analytic (standards-based) marking. There was a moderate to low significant correlation between the teacher's score and the pairs judging score ($r=0.461$, $p<0.01$) and the teacher's semester mark ($r=0.39$, $p<0.01$). The rank of the analytic scores was moderately strongly and significantly correlated with scores from the pairs judging ($r=0.74$, $p<0.01$). The ranking from the individual analytical assessors tended to be correlated low to moderate with the ranking from the pairs judging ($r=0.21$, $p<0.05$ and $r=0.62$, $p<0.1$). For some students, there were substantial differences in the scores awarded by the different methods of marking and in the overall ranking in the population.

Conclusions

The research project demonstrated that digital technologies present considerable potential for assessment of senior secondary courses with a significant performance dimension. Performance in subjects such as Physical Education and Engineering can viably be assessed utilising an integrated task that encompasses the digital capture of student performance through recording practical field activities, web cam commentary, self-reflection, sketching and annotation, and text.

The tasks achieved a notable degree of integration between theoretical and practical aspects of the courses. The potential for this to translate into changed pedagogical practices for the teachers, resulting in an elimination of the theory-practice divide was evident. The tasks could readily be adapted for a number of alternative contexts within Physical Education and Engineering, or in other subjects with access to various practical and computer facilities.

It was also recognised that other aspects of course content could readily be addressed using the techniques adopted in this study: computer-based question-response format, digital video clips as a reference point for questions, idea development through sketch annotation and narratives of design decisions through the web cam.

Results arising from both the analytical and pairs comparison marking pointed to a need for further research to address reliability, and further investigation of the data generated in this project to identify factors contributing to inconsistency in marking. It seems that the analytical and the comparative pairs marking, despite marking to the same criteria (though one is itemised and one holistic), are making a different type of judgement, and possibly the mental integration (Crisp, 2010) involved in making holisitic judgements is a more appropriate form for assessing integrated performance (Penney, Gillespie, Jones, Newhouse, & Cambell, 2011).

The overall task format and the technologies employed were well received by students and teachers working in a range of schools and using varied physical activity and engineering contexts. The additional positive response from assessors and time efficiencies noted in undertaking comparative pairs assessment, and the reliability of marking using comparative pairs across the three years of the project, all lend support to further trialling and research using this method of assessment. The project has also provided an important foundation for ongoing development of assessment by teachers in WA who are working with a range of senior secondary school subjects. Aspects of this research are certainly applicable to courses other than those incorporated in our study and, furthermore, to education contexts in many countries. The theory-practice nexus that this research has sought to engage with is a common feature in a range of subjects, and the opportunity to help overcome this through the digital collection of performance evidence is increasingly available. Similar research is being conducted in New Zealand, Scotland, Ireland, Israel, Singapore and continues in Australia and England.

Contact details for correspondence:
pj.williams@waikato.ac.nz
d.penney@waikato.ac.nz

References

Arnold, P. J. (1979). *Meaning in movement, sport and physical education*. London, England: Heinemann.

Arnold, P. J. (1988). *Education, movement and the curriculum: A philosophic inquiry*. London, England: Falmer Press.

Barnes, M., Clarke, D., & Stephens, M. (2000). Assessment: The engine of systemic curricular reform? *Journal of Curriculum Studies, 32*(5):623–650.

Brown, T., & Penney, D. (2011, June). *Learning 'in', 'through' and 'about' movement in senior physical education? The new Victorian Certificate of Education Physical Education study design*. Paper presented at the AIESEP Moving People, Moving Forward International Conference, University of Limerick, Ireland.

Carbol, B. (2007) *Transition to online testing*. British Columbia, Canada: Society for the Advancement of Excellence in Education.

Crisp, V. (2010). Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education, 36*(1), 1–21.

Curriculum Council of Western Australia. (2002). *Our youth, our future. Post-compulsory education review*. Perth, WA, Australia: Author.

Curriculum Council of Western Australia. (2007). *Engineering studies*. Perth, WA, Australia: Author.

Curriculum Council of Western Australia. (2010). *Physical education studies*. Perth, WA, Australia: Author.

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

Hay, P. (2006). Assessment for learning in physical education. In D. Kirk, D. Macdonald, & M. O'Sullivan (Eds.), *International handbook of research in physical education* (pp. 312–325). London, England: Sage.

Kimbell, R., Wheeler, T., Miller, A., & Pollitt, A. (2007). *e-scape: e-solutions for creative assessment in portfolio environments*. London, England: 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, May). *Assessment to fit for purpose*. Paper presented at the 32nd Annual Conference of the International Association for Educational Assessment. Singapore.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Norway tests laptop exams. (2009, May 1). *BBC News*. Retrieved from http://news.bbc. co.uk/2/hi/8027300.stm

Penney, D., Gillespie, L., Jones, A., Newhouse, P. & Cambell, A. (2011). Assessment in senior secondary physical education. Questions of judgement. In G. Dodd (Ed.), *Edited proceedings of the 27th ACHPER International Conference* (pp. 103–110). Adelaide, SA, Australia: ACHPER. Retrieved from http://www.achper.org.au/conferences-events/2011-conference-proceedings

Pollitt, A. (2004, June). *Let's stop marking exams*. Paper presented at the International Association for Educational Assessment Conference, Philadelphia, PA.

Ridgway, J., McCusker, S., & Pead, D. (2006). *Report 10: Literature review of e-assessment:* Bristol, England: Futurelab.

Thorburn, M. (2007). 'Achieving conceptual and curriculum coherence in high-stakes school examinations in physical education'. *Physical Education and Sport Pedagogy, 12*(2), 163–184.

Wiegers, J. (2010, August). *E-assessment in The Netherlands, innovaitons for the 21st century. Developments since 2000 concerning computer based and computer assisted national examinations for Dutch secondary education.* Paper presented at the 36th International Association for Educational Assessment. Bangkok, Thailand. Retrieved from http://www.iaea2010.com/fullpaper/213.pdf