IJTDE special edition 2012, 22(2)

Investigating the feasibility of using digital representations of work for performance assessment in Engineering.

P John Williams

Abstract

This paper reports on the results of a three-year study conducted at the Centre for Schooling and Learning Technologies (CSaLT) at Edith Cowan University in collaboration with the Curriculum Council of Western Australia which concerns the potential to use digital technologies to represent the output from assessment tasks in the senior secondary course, Engineering Studies. The general aim of this study is 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 was the need to provide students with assessment opportunities in new courses, that are on one hand authentic, where many outcomes do not lend themselves to being assessed using pen and paper over a three hour period, while on the other hand being 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 outcomes without a huge increase in the cost of assessment. 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, video and voice. The exam outputs were uploaded to a online repository. The students' work was marked by external assessors using a standards based rubric that allowed the students work to be ranked though Rasch Modelling.

Keywords: *Digital assessment, Pairs Marking, Computer Managed Learning, High stakes assessment, Online learning, elearning*

Introduction and Background

This research occurred in schools in Western Australia between 2008 and 2010. The study set out to investigate the use of digital forms of assessment in the upper secondary school Engineering Studies course.   It built on concerns that the assessment of student achievement should, in many areas of the curriculum, include practical performance and that this will only occur in high-stakes contexts if the assessment can be shown to validly and reliably measure the performance and be manageable in terms of cost and school environment. The assessment in this project is summative in nature in that it is principally designed to determine the achievement of a student at the end of a learning sequence rather than inform the planning of that sequence for the student.

The study specifically addressed a critical problem for the school system in Western Australia, which also has national and international significance. At the same time the study advanced the knowledge base concerning the assessment of practical performance by developing techniques to represent practical performance in digital forms, to collate these in online repositories, and to judge their quality using both a standards-based marking method and trialling a comparative pairs judging method.

Significance and Rationale

From the 1990's, 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 have provided the opportunity to modify teaching and assessment.

In Western Australia a critical recent concern was the development of high-stakes senior secondary courses such as Engineering, in which at least some of the intended learning outcomes were not able to be adequately assessed using paper-based methods and a three-hour pen and paper exam as the external summative assessment for the course. Therefore it was important that a range of forms of assessment were considered along with the potential for digital technologies to support these alternative forms.

There is a critical need for research into the use of digital technologies to capture student performance on complex tasks for the purposes of summative assessment that are feasible within the constraints of school contexts. This study investigated digital forms of assessment for the Engineering course that aimed for high levels of reliability and manageability, and which were capable of being scaled-up for state-wide implementation in a cost effective manner.

Many educational researchers argue that traditional assessment fails to assess learning processes and higher-order thinking skills, and go on to explain how digital technologies may address this problem (Lane, 2004; Lin & Dwyer, 2006). This argument centres around the validity of the assessment in terms of the intended learning outcomes, where there is a need to improve the criterion-related validity, construct validity and consequential validity of high-stakes assessment (McGaw, 2006). This aligns with Biggs (1999) notion of constructive alignment between the learning activities and assessment asks. Further, in a course such as Engineering, students learn with technologies and this dictates that students should be assessed making use of those technologies. Lin and Dwyer (2006) argue 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) but suggest that this is seldom done due to "technical complexity and logistical problems" (p.28). Dede (2003) suggests that "the fundamental barriers to employing these technologies effectively for learning are not technical or economic, but psychological, organizational, political and cultural" (p.9).

Apart from the lack of validity of traditional paper-based assessment methods, another compelling rationale to consider the efficacy of performance assessment is that teachers tend to teach to the summative assessment (Lane, 2004; Ridgway, McCusker, & Pead, 2006). McGaw (2006) discussed this in the light of changes in the needs of the society, advances in psychometric methods, and improvements in digital technologies, and believed that this was critical due to the influence of assessment on the judgements of teachers and students concerning the curriculum with "risk that excessive attention will be given to those aspects of the curriculum that are assessed" and that "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). A concern underpinning the argument for computer-based assessment methods to replace traditional paper-and-pencil methods was presented by the American National Academy of Sciences (Garmire & Pearson, 2006). They argue that assessing many performance dimensions is too difficult on paper and too expensive using "hands-on laboratory exercises" (p. 161) while computer-based assessment has the potential to increase "flexibility, authenticity, efficiency, and accuracy" but must be subject to "defensible standards" (p. 162) such as the Standards for Educational and Psychological Testing (American Educational Research Association et. al., 1999).

McGaw (2006) also believes that without change to the main high-stakes assessment strategies currently employed there is a reduced likelihood that productive use will be made of formative assessment. He is not alone in this concern, for example, Ridgway et. al. (2006, p. 39) states that, "There is a danger that considerations of cost and ease of assessment will lead to the introduction of 'cheap' assessment systems which prove to be very expensive in terms of the damage they do to students' educational experiences." Therefore, from both a consideration of the need to improve the validity of the assessment of student practical performance, and the likely negative impact on teaching (through not adequately assessing this performance) there is a strong rationale for exploring alternative methods of assessment.

Problem and Research Question

The general aim of the study was to explore the potential of digital forms for external assessment for the senior secondary Engineering course in Western Australia. The problem being addressed was the need to provide students with assessment opportunities in the Engineering course, that are on one hand authentic, where many outcomes do not lend themselves to being assessed using pen and paper over a three hour period, while on the other hand capable of reliable assessment by external examiners. That is, the external assessment for a course needs to accurately and reliably assess the outcomes without a huge increase in the cost of assessment. A significant cost issue in Western Australia derives from the small population dispersed over a large area, making it quite expensive to transport paper portfolios from schools to examiners.

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 Engineering?

This study addressed this question by considering a number of subsidiary questions.

1. What is the feasibility of the digital form of assessment in terms of four critical dimensions: technical, pedagogic, manageability, and functional?
2. Does the paired comparison judgments method deliver reliable results when applied to student practical performance in Engineering?

Performance in Engineering implies a capability related to designing solutions to problems using engineering concepts and knowledge. So students investigate, research, design and make products to solve engineering problems. Paradoxically, the use of paper based exams for assessment calls into question the authenticity of these exams in which the form of assessment does not align with the stated aims of the course, namely to

> ”provide 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” (Curriculum Council, 2007, p1)

The assessment does not align with the prevailing pedagogy, either, which is workshop based with students testing their ideas in a practical engineering context.

There are numerous examples of the use of digital technologies in assessment; however, their use in high-stakes school-level performance assessment is relatively rare, no doubt due to a range of 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. To some extent this is due to a lack of understanding or knowledge, but largely it indicates the need for compelling research findings. Such research needs to start with an understanding of the nature and processes of assessment, and thus the present study not only commenced with a review of the literature but also embedded this understanding within the design for the study, as presented later.

Literature Review

While it will be assumed that the basic constructs within the field of assessment are known and apply perhaps it is useful to be reminded of this through a statement of the three pillars that Barrett (2005) suggests provide the foundation for every assessment:

2. A model of how students represent knowledge and develop competence in a content domain.
3. Tasks or situations that allow one to observe students’ performance.
4. An interpretation method for drawing inferences from performance evidence.

Assessment of Practical Performance

Research in, and the call to investigate “performance-and-product assessment” is not new as pointed out by Messick (1994, p. 14), tracing back at least to the 1960s. A research paper emanating from the *Assessment and Teaching of 21st Century Skills* project focuses on performance in practice; it lays out a clear call to action, arguing that changes are required in high stakes assessments before needed change will occur in schools:

> Reform is particularly needed in education assessment, how it is that education and society more generally measure the competencies and skills that are needed for productive, creative workers and citizens. … more often than not, accountability efforts have measured what is easiest to measure, rather than what is most important. ... New assessments are required that measure these skills ... To measure these skills and provide the needed information, assessments should engage students in the use of technological tools and digital resources and the application of a deep understanding of subject knowledge to solve complex, real world tasks and create new ideas, content, and knowledge (Cisco et al., 2009, p. 1).

Lesgold (2009) echoes this in his validity-based argument for performance assessment but he also recognises the need for reliability, comparability and fairness. Whereas he calls into question the existence of a shared understanding among the general public on what is wanted out of schools, and how this may have changed with changes in society, he argues that these must complement changes to assessment to include 21st century skills in which students respond to tasks representing complex performances, supported by appropriate tools, and with the results judged by experts. The literature clearly depicts the assessment of student performance as critically important but fundamentally difficult, with many unanswered questions that require research.

Methods of Marking

Task assessment is what is commonly referred to as 'marking'. Once students have completed the assessment task the output needs to be judged by some method to determine a score, grade or ranking. Three methods of marking are considered here: 'traditional' true score marking, judgements using standards-based frameworks, and comparative pairs judgements.

Globally, interest in performance assessment has increased over the past decade, with the increasing use of standards-referenced curricula and a focus on educational accountability. Standards-referenced curricula typically define student achievement in terms of what students understand, believe or can do; whereas educational accountability requires that this be measured very accurately or reliably. The issue of the reliability of performance assessment primarily concerns 'marking', with the traditional approach for summative assessment being to, as Pollitt (2004) puts it, sum scores on "micro-judgements" (p. 5) (sometimes called true score marking or cumulative marking). He explains that this approach is likely to generate scores with low reliability for the measurement of "performance or ability" (p. 5). Typically the primary requirement is to provide a ranking of students and therefore, he argues, comparisons between performances using more holistic judgements and Rasch modelling will not only provide this but also a reliable interval scale. The results of implementing a comparative pairs approach to marking that he helped implement for the e-Scape project attested to the saliency of his argument with very positive results (Kimbell, Wheeler, Miller, & Pollitt, 2007). This approach to marking requires assessors to select a 'winner' between the work of a pair of students, and repeat this process many times for many pairs with the results being analysed using a Rasch model for dichotomous data.

Standards-reference or based frameworks and rubrics have been used for many years by teachers in Western Australia and other localities to mark student work but have less often been used for summative high-stakes marking. This involves the definition of standards of achievement against which to compare the work of students. Typically this is operationalised for a particular assessment task through a rubric that describes these standards according to components of the task. The results may be represented as a set of levels of achievement or may be combined by converting these to numbers and adding them. However, using Rasch Modelling they may be combined to create an interval scale score. This paper will refer to this approach as *analytical marking*.

Comparative pairs marking involves a number of assessors making judgements on achievement through comparing each student's work with that of other students, considering a pair of students at a time and indicating the better of the two. This is sometimes referred to as pair-wise comparisons or cumulative comparisons.

Digital Representations of Assessment Tasks

In order to judge student performance, that performance needs to either be viewed or represented in some form. This may involve the assessor viewing a student performing or viewing the results of a student performing. Most often the latter occurs because this is either more appropriate or more cost-effective. However, where the forms of student performance can be recorded in digital representations using video, audio, photographic or scanned documents, or student work that is created in digital format using computer software, the work can be made available to assessors relatively easily and cheaply using digital repositories and networked computer workstations, thereby overcoming the Western Australian distance problem.

A ground-breaking project aimed at assessing performance, titled e-Scape, was conducted by the Technology Education Research Unit (TERU) at Goldsmiths College, University of London (Kimbell, Wheeler, Miller, & Pollitt, 2007). This project built upon many years of work on improving assessment in the design and technology curriculum (Kimbell, 2004). The e-Scape approach represents student work entirely in digital form, collating this work using an online repository, and marks it using a comparative pairs judgements technique. Their results have been encouraging with student work being assessed on an interval scale with a reliability coefficient of 0.93. Kimbell developed a feasibility framework consisting of four dimensions useful for evaluating digital forms of assessment. This framework is described in Table 1.

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.

Table 1. Feasibility 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 presentation 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.

While each of these innovations is not new in themselves, their combination applied at the secondary level of education is new. Apart from Kimbell's (2007) work at the University of London there was no known precedent.

The digital representations of student performance were combined within an online repository. The use of online repositories for student work output is increasingly common, often referred to as online portfolios, with many products available to facilitate their creation and access (Richardson & Ward, 2005). The key feature is that the portfolios can be accessed from anywhere and thus markers from different jurisdictions can be involved, enhancing consistency of standards.

The paired comparison judgement method of marking, involving Rasch Modelling, was implemented using holistic and component criteria judgements. 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, and he does suggest that further research is required to determine the appropriateness 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 first year of the study was a 'proof of concept' stage to explore the feasibility of the digitally-based format for external assessment. The 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 while in the third year this was scaled up to be implemented in a larger sample of representative schools.

The study was evaluative in nature set within an ethnographic framework in that activity was 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. Therefore, this project employed an ethnographic action research evaluation methodology using interpretive techniques involving the collection of both qualitative and quantitative data. The study drew on the traditions of interpretive research but also sought to employ, where appropriate, the quantitative methods of more traditional positivist research. That is, quantitative measures concerning achievement, costs and resource use were used but need to be interpreted within the context of the learning environment and in which the measures occurred.

The research design can be described as participative action research with participants contributing to development through evaluative cycles. As such this required an analysis of the perspectives of the key groups of participants (teachers, assessors, students) with data collected from each group. This approach allowed for refinement and further development of findings based on multiple instances of the same phenomenon under different conditions (Willig, 2001). Therefore, this study largely employed an ethnographic action research evaluation methodology using interpretive techniques involving the collection of both qualitative and quantitative data.

The study sought to involve a broad range of schools and teachers, some of which had been involved in this study in all three years, others for just one year. In the third year the study involved 8 teachers and their 8 upper secondary classes containing 94 students. Data from all sources on each class was combined to create a series of case studies structured to highlight design and implementation features. For all cases the course was Engineering Units 1 or 2, Year 11 students, specializing in the range of options available: Mechanical, Electrical or Control.

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

The quantitative and qualitative data for each class were compiled into case studies that were sent to the teacher(s) involved for validation and some interpretation. Triangulation of data types and sources enhanced the credibility of findings. The outcomes from each case study were then summarised according to the dimensions of feasibility so that the data could be analysed using a constant comparative approach looking for themes, trends and developing rich descriptive accounts (Patton, 1990). Data were coded according to emergent themes. Themes were constantly compared with emergent categories to establish a best fit with the data set.

On the basis of the analysis of the implementation of the assessment tasks in the second year of the study and on the situation analysis conducted in the first year, the working team refined the assessment task for the final year.

Marking criteria were developed from the assessment task and the course syllabus outcomes and content and presented in 'marking key' form or rubric form for the analytical marking. One holistic criterion was distilled from the standards-referenced criteria for use with the comparative-pairs marking.

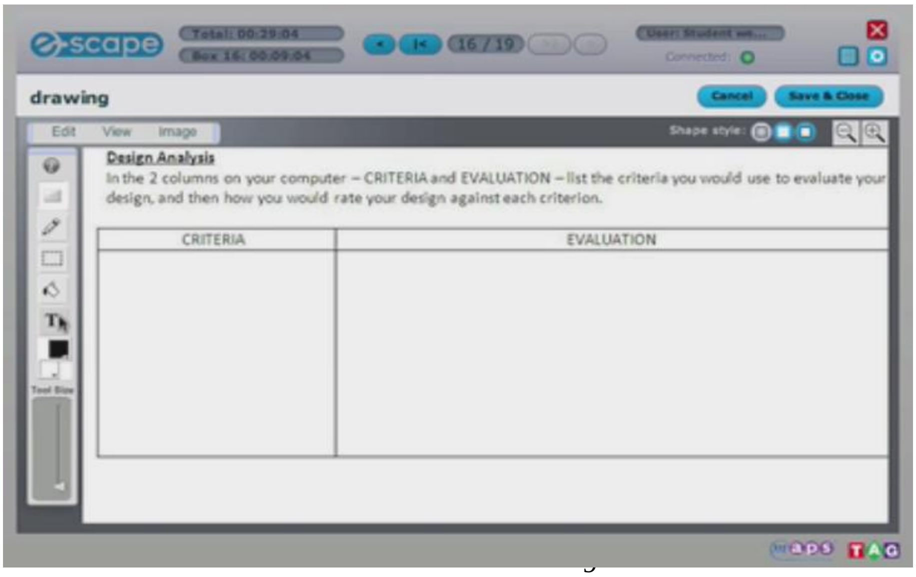
The Assessment Task

A set of tasks was provided as a scaffold to responding to a design brief. The design 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 the sea water. 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 the materials data table. The expectation was that the students would progress their design over the four iterations. They used the webcam to photograph their sketches, and then annotated them online; they 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.

Two elements of the examination were changed for the third year. The first was related to the collaboration which occurred amongst the students doing the exam. Their initial sketches were viewed by a fellow student for critical feedback, the logic being that critiquing is an important element of designing and the examiners could then assess this ability to critique. It would also provide the opportunity for students to respond to their peers' comments and develop a more effective design as a result. This collaborative element was removed because it required having an even number of students in the class, and the examining authority felt uncomfortable with the notion of students sharing ideas during an examination.

The second element of the examination that was changed in the final year of the project was the material modelling that students engaged in as part of their process of designing a solution to the problem. In the second year, the research team supplied each class with modelling materials which they could use to develop their design ideas at a specified time within the design process. This was quite time consuming, and when considering scaling the examination to be delivered across the state, may be problematic. So this aspect of the design was replaced with an alternative form of stimulation, namely a data table which provided students with technical data for the range of materials they had available for their design upon which they could base a revised and improved solution to the problem.

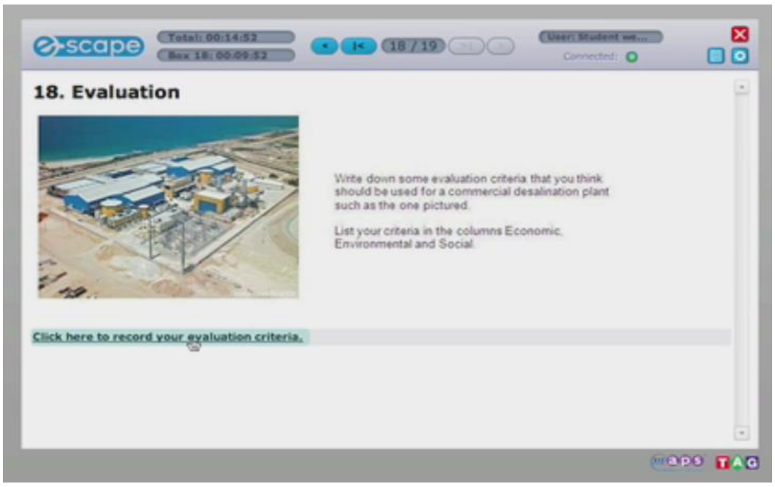
The e-scape exam management system was used as the tool to design and present the task to the students. This is a program that enables the design of a portfolio template into which students can input a range of forms of data – text, graphics, voice and video (Figure 1). Stimulation material and instructions were included in the portfolio and each task (or portfolio 'page') was timed. The students were required to do some sketching of their ideas on paper, and then they took a picture of their sketch to include in their e-scape portfolio. A paper template was prepared for this purpose, folded to promote the sequence of activities required, and printed on 2 sides of an A3 sheet. The assessment task output for the students was collected using the e-scape online exam management system that uploads into the MAPS portfolio server.



Design Analysis

In the 2 columns on your computer – CRITERIA and EVALUATION – list the criteria you would use to evaluate your design, and then how you would rate your design against each criterion.

CRITERIA	EVALUATION



18. Evaluation

Write down some evaluation criteria that you think should be used for a commercial desalination plant such as the one pictured.

List your criteria in the columns Economic, Environmental and Social.

Click here to record your evaluation criteria.

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

Technologies

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

1. ***Intranet***. In this scenario all the students were issued with ASUS EeePC mini computers that 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. The students' examination portfolios were automatically uploaded to the research facilitator's computer at the conclusion of the examination. They then had to be later uploaded to the examination server in preparation for assessment. The examination of one group in one school was conducted this way.
2. ***Live***. In this scenario the students were working on the school computers and logged on live through the internet to the examination server. The research facilitator was also logged on as the examination manager and was then able to monitor each logged-in students progress throughout the examination tasks and move the students on to each task when the time expired. The students' examination portfolios were automatically uploaded to the server at the conclusion of the examination. The examination of three groups in three schools was conducted this way.
3. ***USB***. In this scenario the examination was accessed by the students through logging on to a USB drive that was inserted into a school computer. This did not enable centralized control by the research facilitator of the examination administration, but as students were able to progress through the exam at their own rate, central control was not necessary. The USB's were collected at the end of the examination period and later uploaded to the examination server for access by the assessors.

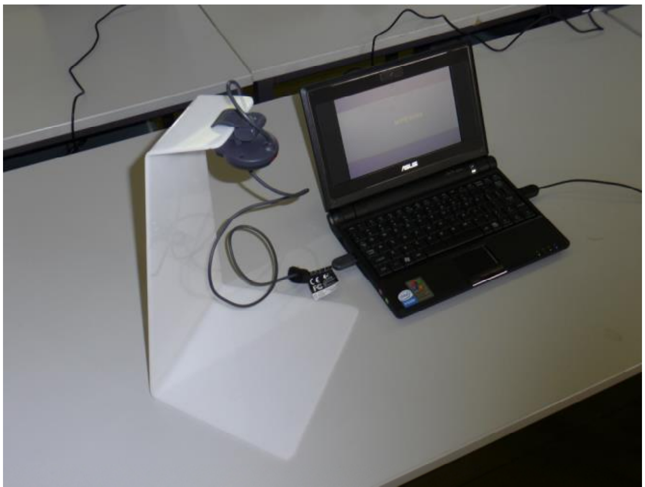
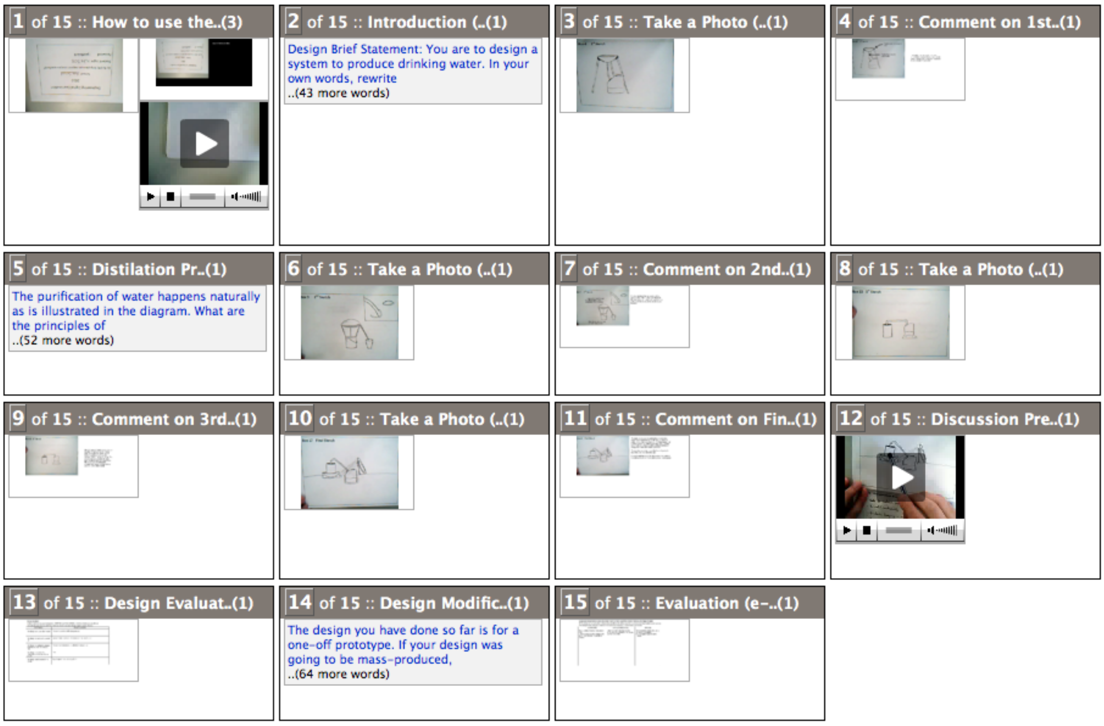


Figure 2. An ASUS EeePC computer with stand to hold external USB camera.

Marking of Student Work

The outputs for the 94 students were uploaded to an online repository. All marking of work was done using online digital tools that accessed the digital repository. The two external assessors marked the work for each student from the examination using an analytical marking method involving the use of rubrics. Teachers were requested to mark the examination and award a semester mark for the student. Later, all work was marked again using a comparative pairs method. Figure 3 shows how the student portfolio was presented to the examiner, and the screen for recording the comparative pairs judgement.

There was a moderately significant correlation between the two assessors using the analytical method, with correlation coefficients of 0.53 and 0.49 (p<0.01) for both scores and ranking of students. This would tend to indicate that the scores were reasonably reliable and that the marking guides maybe could have been more explicit.



1 of 15 :: How to use the..(3)

2 of 15 :: Introduction (..(1)

Design Brief Statement: You are to design a system to produce drinking water. In your own words, rewrite
..(43 more words)

3 of 15 :: Take a Photo (..(1)

4 of 15 :: Comment on 1st..(1)

5 of 15 :: Distilation Pr..(1)

The purification of water happens naturally as is illustrated in the diagram. What are the principles of
..(52 more words)

6 of 15 :: Take a Photo (..(1)

7 of 15 :: Comment on 2nd..(1)

8 of 15 :: Take a Photo (..(1)

9 of 15 :: Comment on 3rd..(1)

10 of 15 :: Take a Photo (..(1)

11 of 15 :: Comment on Fin..(1)

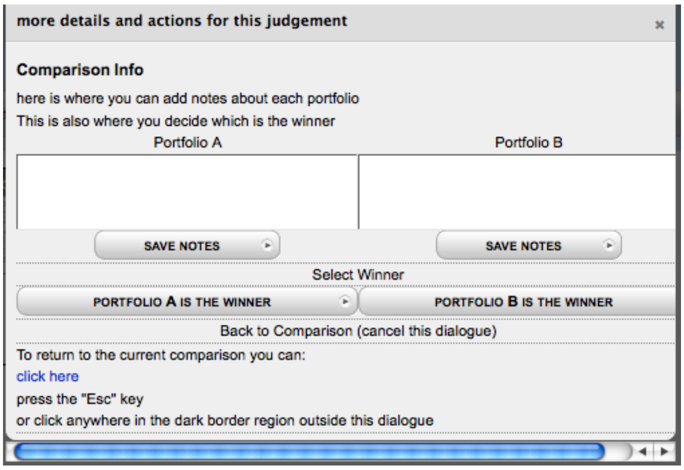
12 of 15 :: Discussion Pre..(1)

13 of 15 :: Design Evaluat..(1)

14 of 15 :: Design Modific..(1)

The design you have done so far is for a one-off prototype. If your design was going to be mass-produced,
..(64 more words)

15 of 15 :: Evaluation (e-..(1)



more details and actions for this judgement

Comparison Info

here is where you can add notes about each portfolio

This is also where you decide which is the winner

Portfolio A
Portfolio B
SAVE NOTES
SAVE NOTES
Select Winner
PORTFOLIO A IS THE WINNER
PORTFOLIO B IS THE WINNER
Back to Comparison (cancel this dialogue)

To return to the current comparison you can:

click here

press the "Esc" key

or click anywhere in the dark border region outside this dialogue

Figure 3. Screen displays of the student portfolio and the tool used for comparative pairs judging.

There were moderate to low significant correlations between the average score of the external assessors and the scores awarded by the teacher for the examination and for the semester (r=0.64 and 0.46, p<0.01). Correlations of ranking (the order in which the marks awarded placed the student in the population) were similar. This suggests that, in spite of the differences in content assessed, student competence was recognised consistently by both assessors and teachers.

Comparative pairs judging involved the assessment of the exam output for 94 students. The holistic criteria used for assessment was: is 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?

Seventeen assessors completed 473 comparisons between the exam outputs of pairs of students using the *Pairs Engine (see the article by Pollitt)*. The pairs to be judged were dynamically generated by the software. It had been decided to stop marking once the Cronbach Alpha Reliability Coefficient was 0.9 or better. This occurred after the 10th round of marking when this coefficient was 0.927. Thus a reliable set of scores was generated. 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. There was a moderate significant correlation between the teacher's score and the pairs judging score (r=0.39 p<0.01) and the teacher's semester mark (r=0.36 p<0.01).

For some students there were substantial differences not only in the scores awarded by the different methods of marking but also the overall ranking in the population. Each student with a difference in ranking between the two methods of marking of more than 45 was considered in depth by reviewing the scores allocated, the comments of assessors and the nature of the work output. There were 10 such students. No clear pattern emerged from this analysis to explain the differences in rank that applied only to these 10 students.

Findings

The findings are summarised under the categories of the feasibility framework (Table 1).

Manageability

- The two schools involved in the previous cycle in 2009 had no management problems. The remaining six schools had technical set-up issues related to using their computers that were resolved prior to the examination through discussions with technicians and testing of the software. Due to difficulties with Internet access speed, four of these six schools needed alternative solutions to access the exam. The solutions were to use an Intranet wireless server (one school) and to use a USB drive version of the exam in three schools including one remote and one country school.
- Ideal rooms were computer labs with plenty of desk space. This was found necessary for the sketching and the recording of the video sequence. If the students were seated too close to each other, the video discussion was particularly disruptive. In some rooms the students were placed at every other computer to allow for more room between them.

Internet Live

- This method was the easiest to manage, however as more students began to access the Internet, speed was reduced and a number of students found it frustrating uploading pictures and completing the video component. A disadvantage was that if one student did not complete a section of the examination during the allocated time (for eg uploading a picture), then the whole class had to be redirected to that section again.

Intranet

- Needed to test that the intranet wireless server worked within the school environment.
- A number of students commented that the Notebook screen and keyboard was too small for this type of exam.

USB drive

- One advantage was that the students were able to work through the exam at their own rate.
- Once set up this provided minimal running technical issues.
- The school computer operating system needed to support the executable files that were on the USB drive. This was overcome in all cases.
- Problems were encountered in uploading the students' work to the server. This should have involved simply loading the USB student work through a live Internet connection.

Technical

- In all schools there were few technical impediments to the implementation of the assessment task.
- Technical issues during the exam related to the web-cams and uploading pictures.
- In most cases dependence on school infrastructure was reliable.
- Hardware was generally reliable – some web-cam uploads were very slow causing students to try and force the system resulting in some freezing of computers and meant the student needed to re-login.
- A few students needed to relocate to another computer due to technical problems.

Internet Live

- As students began to access the Internet in the exam, as well as other classes in the school, the uploading and downloading of student work slowed and this caused frustration and freezing of some computers.

Intranet

- The process of uploading the class work to the server was done with few technical difficulties.

USB drive

- The process of uploading each individual USB presented both a time issue and a number of significant technical issues.

Functional

- All teachers indicated that the structure and process of the exam was good. Some commented that technical problems distracted from the student exam experience. One teacher felt that a marking key was needed and that seeing other student work and hearing their video presentation was a problem.
- In all cases most students readily perceived the assessment task to be authentic, engaging and meaningful for their course.
- It was generally thought to be a well-structured and sequenced task that generally provided the students with a good framework to tackle the task, although it was obvious some students ignored the stimulus material provided and only made slight modifications to their initial choice of solution.
- The task was structured to permit a good range of levels of achievement to be demonstrated, but one teacher felt it could have been broader and more ‘engineering like’.
- While correlations between assessors and teachers were generally low and not consistent, overall correlations were significant, and strong between analytic and pairs markers.
- The exam allowed discrimination between students of varying ability.
- For analytical marking the inter-rater reliability was significant but only moderate ($r=0.53$, $p<0.01$) as was the correlation with teacher scores ($r=0.64$, $p<0.01$).
- Although Rasch analysis of the results of analytical marking yielded an initial Cronbach Alpha index of 0.890 and Pearson Separation index of 0.897, some criteria showed thresholds disorder though in general they fitted the model. Therefore, some items were re-scored by collapsing the relevant categories that showed better results in terms of threshold order.
- Comparative pairs judging results achieved a Cronbach-Alpha reliability coefficient of 0.927 after 11 rounds of judging. Scores generated were considered to be reliable with the system, only finding 2.75% of judgements highly suspect.
- Differences between the scores from the two analytical markers did not correlate with differences between the analytical scores and the comparative judgement results. Therefore it is likely that different factors contribute to differences between the two methods as opposed to between two individual assessors.

Pedagogic

- Typically students liked doing an assessment task in this form.

- It was thought to be a little repetitive in parts and the overall task could be more "strenuous" to develop higher order thinking. The repetitive aspect concerned the need for four sketches over the full task.
- The assessment task matched the typical pedagogy for the course, with the exception of one school where the students had not done much design.
- All students preferred the assessment task than a paper-based exam.
- Most students had done design tasks on a computer before.
- Most teachers believed the examination gave students a fair opportunity to demonstrate their understandings of the course, though a number said they would have liked to see a more 'engineering' type of task.

Conclusions

This study identified relatively few constraints to the use of the digital form of performance exam for the Engineering Studies course implemented in the sample of schools. In all schools a suitable computing room was able to be sourced with only two schools where it was necessary to have two rooms side by side because of computer failures and large numbers of students. Most schools had initial trouble with web cams but these were generally overcome fairly quickly. There were many situations where computers froze however, particularly during the process of uploading and then downloading pictures and this contributed to the frustration of students.

Although some students indicated a little more time would have been useful, most of the students completed the task within the time. The design task presented in the examination was not a typical engineering task, but the focus of this research was on the procedure of trialling the examination rather than on the content. The methodology used would be equally applicable to a task that students would associate more with the type of engineering they do in school.

Overall the third year of the study found that the benefits of the digital form of performance exam implemented outweighed the constraints. Differences were revealed between schools where the focus had been more on the specialization of the Engineering course than on the more design oriented core content. The assessment aligned well with pedagogy of Engineering Studies, and most students enjoyed the practical design nature of the examination task.

The research demonstrated that a computer-based performance exam could be constructed for the core content of the Engineering Studies and that this could be readily implemented in a large range of schools offering the course. However, in a number of ways minor improvements could be made to the structure, content and implementation of the exam, which may also improve the reliability of analytical marking.

The vast majority the students were very engaged and enthusiastic about undertaking the exam. The structure of exam is perceived by some students to be over-scaffolded with four iterations of design solutions required in response to stimulus material. This could be overcome with a more comprehensive introduction to the examination and practice by way of mock exams prior to sitting the exam.

It was recommended that the implementation of the exam should initially use USB flash drives for delivery and collection of student work with an online upload at the end of the exam as a backup. In some schools the live upload of examination material worked well, but not all schools have this capacity. It is highly likely that the exam can be adequately implemented in all schools involved provided that:

- the students' regular computer laboratory can be used;
- there are about 10% spare workstations;
- a regular IT support person is on standby;
- the invigilator is equivalent to a teacher in the course in terms of understanding of the technology (e.g. each Engineering teacher could invigilate at another school);
- an audit is completed to check that software is available for web-cams and for running the USB codes.

The marking of the exam should initially use a criteria-based analytical method. Given the considerable difference in ranking between the two methods of marking that was found in this third year it is recommended that analytical marking with Rasch modelling is used initially rather than the comparative pairs method. This method of marking currently has public confidence and has been shown through this research to generate scores that are adequately reliable, which would increase with more rigorous examiner training.

In conclusion it was recommended that an exam similar to that implemented in the third year be instituted as quickly as possible, as a component of the external summative assessment in the course, and that during the first few years further research be conducted to investigate migrating the delivery and collection of exam materials using an online system. However, at this point in time this research has not found such a system with adequate reliability and therefore recommends that the exam be implemented using local storage technology such as USB drives, which could incorporate a live upload to the server which would enact in those schools where this facility is available. In this case the USB drive would become a backup.

References

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Cisco, et. al. (2009). Transforming Education: *Assessing and Teaching 21st Century Skills.* Assessment and Teaching of 21st Century Skills. Retrieved September 8, 2011 from http://atc21s.org/.

Barrett, H. C. (2005). *Researching Electronic Portfolios and Learner Engagement*: The REFLECT Initiative.

Biggs, J.B. (2003). *Teaching for quality learning at university.* Buckingham: The Open University Press.

Burns, R. B. (1996). *Introduction to research methods*. South Melbourne, Australia: Addison Wesley Longman Australia Pty. Limited.

Curriculum Council. (2007). *Engineering Studies Syllabus*. Perth, Western Australia: Curriculum Council.

Campbell, A. (2008). *Performance Enhancement of the Task Assessment Process through the application of an Electronic Performance Support System.* Unpublished Doctorate, Edith Cowan University, Perth.

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

Garmire, E., & Pearson, G. (Eds.). (2006). *Tech Tally: Approaches to Assessing Technological Literacy.* Washington: National Academy Press.

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. (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.

Lesgold, A. (2009). *Better schools for the 21st Century.* Retrieved 9/6/2009, from http://atc21s.basecamphq.com/clients.

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.

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

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. Newbery Park, California: SAGE Publications, Inc.

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

Richardson, H. C., & Ward, R. (2005). *Developing and Implementing a Methodology for Reviewing E-portfolio Products.* Wigan, UK: Joint Information Systems Committee.

Ridgway, J., McCusker, S., & Pead, D. (2006). *Report 10: Literature Review of E-assessment.*: Futurelab.

The British Psychological Society. (2002). *Guidelines for the Development and use of Computer-Based Assessments*. Leicester, UK: The British Psychological Society.

The Council of the International Test Commission. (2005). *International Guidelines on Comptuer-Based and Internet Delivered Testing.* Granada, Spain: The Council of the International Test Commission.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.

Willig, C. (2001). *Introducing qualitative research in psychology adventures in theory and method*. Buckingham: Open University Press.