1-1-2011

Using IT to Assess IT: Towards Greater Authenticity in Summative Performance Assessment

Christopher P. Newhouse
Edith Cowan University

10.1016/j.compedu.2010.08.023
This is an Author's Accepted Manuscript of: Newhouse, C. P. (2011). Using IT to assess IT: Towards greater authenticity in summative performance assessment. Computers and Education, 56(2), 388-402. Available here

Using IT to assess IT: towards greater authenticity in summative performance assessment

Abstract

An applied Information Technology (IT) course that is assessed using pen and paper may sound incongruous but it is symptomatic of the state of high-stakes assessment in jurisdictions such as Western Australia. Whereas technology has permeated most aspects of modern life, including schooling, and more has been demanded of education systems in terms of outcomes and participation, methods of summative assessment have changed little and are seriously out of alignment with curriculum, pedagogy and the needs of individuals and society.

This paper reports on an analysis of some of the data from a component of a study into the feasibility of using digital technologies to achieve greater authenticity in summative performance assessment in the Applied Information Technology (AIT) course in Western Australian secondary schools. In the first phase of the study a sample of 115 students completed a digital portfolio and a computer-based exam that were both externally assessed using online tools and by two methods of marking, with the results analysed using Rasch modelling software. A traditional analytical method and a comparative pairs method of marking were investigated.

The study found that both the digital portfolio and computer-based exam were implemented without significant technical difficulty and were well accepted by the students and teachers. The work output in digital form was readily accessed from an online repository by external markers using a standard web browser. The two methods of marking provided highly reliable scores, with those from the comparative pairs method being the more reliable. A number of questions of validity and manageability were raised and the strengths and weaknesses of the two forms of assessment revealed. It was concluded that it was feasible to implement either form of assessment for high-stakes purposes, with a resulting improvement in alignment and authenticity.

1. Introduction

It is perhaps self evident that what is taught should be assessed and what is taught should reflect the needs of individuals and society. However, most often the reality is that what is taught is what is assessed and what is assessed bears little resemblance to what is needed (Lane, 2004; Ridgway, McCusker, & Pead, 2006). And what is assessed, particularly for high-stakes purposes, is determined by what can readily be represented on paper using a pen, in a short amount of time (Clarke-Midura & Dede, 2010). Most often this is in stark contrast to the stated intentions of the curriculum content and preferred pedagogy and does not match the requirements of future study, work or life activities. That is, present assessment lacks alignment and authenticity, yet it remains the dominant force driving education systems (Clarke-Midura & Dede, 2010).

This problem, while not new, is perceived to be growing in complexity and proportion (e.g. Dede, 2003; Lane, 2004; Lin & Dwyer, 2006; McGaw, 2006). McGaw (2006) noted that the impact of summative assessment on the curriculum was a critical concern with "risk that excessive attention will be given to those aspects of the curriculum that are assessed", that "risk-taking is likely to be suppressed" (p.2) and less likelihood that productive use would be made of formative assessment. He went 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). He is not alone; for example, Ridgway (2006, p. 39) similarly expresses concern that, "considerations of cost and ease of assessment" will have negative consequences for students. 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.

There are many examples of forms of assessment highly aligned and with great authenticity but in general these are considered to be expensive and difficult to manage, which restricts their application to smaller numbers of students and to particular circumstances (Garmire & Pearson, 2006). For example, the final assessment of performance for accrediting pilots, surgeons and even teachers tends towards high levels of authenticity but this is not the case for the assessment of school students in courses with large enrolments. Few would argue against greater authenticity in the assessment of performance, but they may have concerns with the feasibility of measures to do so either in terms of cost-effectiveness or managing validity and reliability factors (McGaw, 2006; Messick, 1994). The objective is to capture valid performance that may be judged in a reliable fashion, all within budgetary constraints. For example, a performance in 'conducting science experiments' may be assessed by an expert observing a student conducting an experiment and making judgements according to pre-determined criteria. While this would be seen to be a highly authentic assessment (i.e. high content and construct validity) it has a high cost that may be considered excessive when applied to thousands of school students; moreover, the resulting judgements could be questioned in terms of reliability and it may be difficult to ensure consistency of conditions for all students. So, as Garmire and Pearson (2006) point out, whereas assessing many performance dimensions is too difficult on paper it is too expensive and unreliable using "hands-on laboratory exercises" (p. 161).

The quest for authenticity in assessment is quite complex; in addition to assessing practical skills, it concerns assessing higher-order thinking and learning process skills demonstrated through complex performance and as such, many educational researchers would argue, traditional paper-based assessment does very poorly (Lane, 2004; Lin & Dwyer, 2006). The concern of these researchers centres on the validity of such 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 (for definitions refer to McGaw, 2006). Kozma (2009) claims that tasks in the "outside world" require cross-discipline knowledge, relate to complex ill-structured problems, and are completed collaboratively using a wide range of technological tools to meet needs and standards. These characteristics are at odds with traditional pen-and-paper approaches to assessment in schools where problems are necessarily simplified and structured and must draw on narrow sources of information. Therefore, there is a need to consider alternative approaches with alternative technologies, in particular for the representation of student performance on complex tasks.

In applied IT courses digital technologies not only provide pedagogical support, as for many other courses, they are also the context for the content; therefore performance implies capability in using the technologies. It may appear self evident that assessment of such performance would require the use of these technologies; however, paradoxically three-hour paper-based exams are still being used for the Applied Information Technology course in Western Australia – representing an extreme example of the authenticity problem. This form of assessment does not align with the stated aim of the course, which is to provide "opportunities for students to develop knowledge and skills relevant to the use of ICT to meet everyday challenges", nor with the predominant pedagogy that includes students spending most of their time using technologies to create digital products. Furthermore, the focus on theoretical content in a paper-based exam is of limited future value to students pursuing IT-related skills, attitudes and understanding for work and life. There are several ways students could be assessed on their use of digital technologies – typically through portfolio or computer-based exam, each with its strengths and weaknesses. Therefore, the research question becomes: which form of assessment is most feasible for the course under prevailing conditions?

In 2007, with the provision of a new set of high-stakes senior secondary courses, the authenticity problem became critical for schooling in Western Australia with many of these new courses including a major component involveing performance of practical capabilities, in many cases using a variety of technologies. Clearly these skills and knowledge were not conducive to assessment using a three-hour paper-based exam. Therefore alternative forms of assessment had to be devised. To this end, researchers at the Centre for Schooling and Learning Technologies (CSaLT) at Edith Cowan University (ECU) commenced a three-year study with the Curriculum Council of Western Australia to investigate the feasibility of using digital technologies to support assessment tasks in four of these courses: *Applied Information Technology* (AIT), *Engineering Studies*, *Italian* and *Physical Education Studies*. This paper reports on the first year of the study for the component involved with the AIT course. The study sought to use digital technologies for the capture, collation, marking and analysis of student practical performance in AIT. In the first year this involved a sample of 115 students undertaking a digital portfolio and a computer-based exam that were both externally assessed using online tools and by two methods of marking, with the results analysed using Rasch modelling software. The traditional analytical method and a comparative pairs method of marking were applied to test Pollitt's (2004) assertion that the traditional method would generate less reliable scores. The author was the director of the research team that undertook the study.

This paper will start with a brief review of areas of the research literature that underpin the study and provide a theoretical framework. Then the design and method for the study will be introduced, followed by the presentation of some of the results and a discussion of key findings. Finally some conclusions will be drawn that could inform practice and generate further research.

1.1. Digital technologies and performance assessment

Since the 1960s educators have postulated uses for digital technologies in assessment processes, commonly referred to as computer-based assessment, and sometimes applications have found wide acceptance, such as with the automated marking of multiple choice questions and the statistical analysis of assessment scores. As a logical extension of this work many educators have suggested that the assessment authenticity problem may be addressed through the use of digital technologies, and have suggested that the technology may be used to record or represent a performance or to support the marking or analyses processes (Dede, 2003; Lane, 2004; Lin & Dwyer, 2006; McGaw, 2006). The committee for the American National Academy of Sciences cites the use of computer-based adaptive testing, simulations, computer-based games, electronic portfolios, and electronic questionnaires as having potential (Garmire & Pearson, 2006). However, they raise concerns about the use of computer-based assessment methods: whereas they have the potential to increase "flexibility, authenticity, efficiency, and accuracy", they must be subject to "defensible standards" (p. 162), such as the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). The committee identifies a number of topics requiring research, including the use of electronic portfolios that "appear to be excellent tools for documenting and exploring the process of technological design" (p. 170).

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.

1.2. Computer-supported assessment

In order to consider how digital technologies might support assessment, we need firstly to clarify the nature and processes of assessment. Barrett (2007) suggests three pillars that provide the foundation for every assessment:

1. a model of how students represent knowledge and develop competence in a content domain
2. tasks or situations that allow one to observe students' performance
3. an interpretation method for drawing inferences from performance evidence

Digital technologies may be used to support these pillars by providing environments and tools for (1) the representation of knowledge, (2) the recording of evidence to 'observe' performance and (3) the process of interpretation and drawing inferences. Hitherto digital technologies have mainly been used for multiple-choice tests and collating marks, which Lin and Dwyer (2006) describe as very limited routine applications rather than more powerful applications that capture "more complex performances" (p.29), and entail more sophisticated methods of representing knowledge and recording of evidence. The performances Lin and Dwyer are envisaging involve the demonstration of higher-order skills such as decision-making, reflection, reasoning, and problem solving. Whereas here the focus is largely on the first two pillars, it is clear that for high-stakes assessment the third must be considered at the same time; that is, how the evidence of performance can be interpreted or 'marked'.

The majority of the published research in the field of computer-supported assessment relates to higher education (e.g. Brewer, 2004), with little specific to the school sector. However, assessment of secondary student creative work in some areas of the arts has been addressed for some time, although even here Madeja (2004) argues for more use of alternatives to paper-and-pencil testing. There has been some research into the use of portfolios for assessment but most often this was for physical, not digital, portfolios. There has been considerable research into the use of online testing but not involving assessing practical performance – merely replicating paper-and-pen tests in an online environment (e.g. MacCann, 2006).

There has been increasing interest internationally in computer support for assessment with, for example, the recent report from the Joint Research Centre for the European Commission titled, *The Transition to Computer-Based Assessment* (Scheuermann & Bojornsson, 2009). The University of Cambridge Local Examinations Syndicate conducted over 20 projects to explore the impact of new technologies on assessment, including using online simulations in assessing secondary school science investigation skills (Harding, 2006). Other organisations (e.g. Becta, 2006) or groups of researchers (e.g. Ridgway et al., 2006) have reported on exploratory assessment projects, particularly the increasing use of online testing, although rarely for high-stakes assessment and not without some difficulty (Horkay, Bennett, Allen, Kaplan, & Yan, 2006). Most recently a large international commercially supported study has focussed on the assessment of '21st Century skills' – considering a vast array of forms of computer support for assessment or e-assessment (Cisco, Intel, & Microsoft, 2009, p. 1).

Ripley (2009) defines e-assessment as "the use of technology to digitise, make more efficient, redesign or transform assessments and tests". He discusses two 'drivers' of e-assessment: business efficiency and educational transformation. The former leads to "migratory strategies" (i.e. replicating traditional assessment in digital form), whereas the latter leads to "transformational strategies" that change the form and design of assessment. An example he cites of the latter is the ICT skills test

conducted with 14-year olds in the UK in which students complete authentic tasks within a simulated ICT environment. Ripley raises issues that need to be addressed, including: providing accessibility to all students; the need to maintain standards over time; the use of robust, comprehensible and publicly acceptable means of scoring student's work; describing the new skill domains; overcoming technological perceptions of stakeholders (e.g. unreliability of IT systems); and responding to the conceptions of stakeholders about assessment. Kozma (2009) also alludes to similar drivers and raises similar issues that relate to all three pillars cited by Barrett (2007) above, particularly the third – drawing inferences from performance evidence. This leads now to a consideration of the assessment of practical performance.

1.3. Assessment of practical performance

An earlier version of assessment of practical performance, was referred to as "performance-and-product assessment" (Messick, 1994, p. 14), which may be traced back to the 1960s. This terminology meant that the end product of a performance was assessed, or the process of the performance was assessed, or both were assessed. What was assessed depended on the specifics of the situation, which Messick describes as social values that require close attention to the intended and unintended consequences of the assessment through considerations of the purposes of the assessment, the nature of the assessed domain, and "construct theories of pertinent skills and knowledge" (p.14). However, how performance is assessed needs to address the traditional assessment principles of "validity, reliability, comparability and fairness" (p. 14). By its nature, performance assessment tends to address validity but it needs to consider the effect on the other principles when performance is the vehicle of assessment. Often this is stated in terms of replicability and generalisability, and although not restricted to digital forms of performance assessment, these should be its guiding principles (Clarke-Midura & Dede, 2010).

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). 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. In his paper he explains this method of marking and how an interval scale is generated. 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.

A recent research paper (Cisco et al., 2009) emanating from the *Assessment and Teaching of 21st Century Skills* project focusses on performance in practice; it lays out a clear call to action, arguing that changes are required tin 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. He recognises the issues that this would uncover and proposes 'stealth assessment' as an example solution. In stealth assessment students complete a portfolio of performance at school over time that is supervised by the teacher. The testing system then selects one or two "additional performances" to be externally supervised "as a confirmation that the original set was not done with inappropriate coaching" (p. 20). Many such solutions are being envisaged but need to be tested under realistic conditions.

This body of literature clearly depicts the assessment of student performance as critically important but fundamentally difficult, with many unanswered questions that require research. The focus is now narrowed to performance in applied IT courses.

1.4. Using IT to assess IT

The two broad categories of digital forms of assessment considered for the AIT course, digital portfolios and computer-based exams, are now reviewed in more detail.

1.4.1. Digital portfolios

A digital portfolio, or e-portfolio, is an organised collection of items stored in digital files; however, its overall form and structure can vary according to its nature and purpose. For example, an e-portfolio may be a collection of finished products to be used by the student to demonstrate competence in applying for a course or employment. Alternatively, an e-portfolio may be a collection of student work output from a set of processes or aligned with a set of outcomes to demonstrate learning or overall performance. Koretz (1998) defined portfolio assessment, whether digital or physical, in terms of the latter, as the use of a cumulative collection of student work to evaluate performance. He analysed the outcomes of four large-scale high-stakes 'physical' portfolio assessment systems in USA school systems in the 1990s, concluding that overall the programs were resource intensive and did not produce "evidence that the resulting scores provide a valid basis for the specific inferences users base on them…" (p.332). He noted significant improvements from earlier years in the implementation and reliable marking of portfolios, but concluded that portfolio-based assessment was "problematic" (p.309) in terms of manageability. These problems with the validity, reliability and manageability of portfolio performance assessment provide a rationale for researching the feasibility of digital solutions.

Those such as Koretz (1998), Barrett (2007) and Beetham (2005) indicate that the main concerns educational leaders have with the use of digital portfolios for assessment are:

- the authentication of student work given the period of time within which work is completed
- whether digital portfolios are fair to all students in terms of access to information, materials and tools
- whether digital portfolios can be marked reliably given the usually varied types of student work output

Therefore, it is often recommended that the portfolio have a particular structure, with limits on the type, size, time, along with a need for the work to be authenticated by a teacher and the student. Carney (2004) stipulates a set of critical dimensions of variation for digital portfolios, and Barrett (2007) suggests defining characteristics for "Portfolios used for Assessment of Learning". Beetham (2005) points out that whereas in the past e-portfolios had been found to take longer to moderate and mark, they have become more streamlined when part of an "integrated assessment facility". She provides five commercial examples of such systems, and a list of nine "issues relating to the use of e-portfolios for summative assessment" (p. 5), with seven being technical in nature and all but three being readily addressed by the use of a good management system. The remaining issues outlined by Beetham (2007) are:

- acceptability and credibility of data authenticated by awarding bodies
- designing assessment strategies to make effective use of the new tools and systems
- ensuring enhanced outcomes for learners, such as higher motivation, greater choice over evidence, assessment around capabilities and strengths

Beetham (2007) also raises some issues for teachers and learners (p. 16):

- the fit with existing practices and expectations
- access to ICT and the ICT capability of teachers and learners
- acceptability and appropriateness of e-portfolio use by teachers and learners

Arguably most of these issues would not be relevant for AIT because portfolios have been normal practice over many years for school-based assessment. Therefore, provided there has been a good assessment management system the only issue that may not have been addressed currently for AIT is the "Acceptability and credibility of data authenticated by Awarding Bodies" (Beetham, 2007, p. 16).

1.4.2. Computer-based exams

Different types of computer-based exams have been devised where students use computer systems to complete tasks or respond to questions. The simplest form is the answering of multi-choice and short-answer questions on the screen (Siozos, Palaigeorgiou, Triantafyllakos, & Despotakis, 2009), and the most complex the use of various software packages to create digital products (MCEETYA., 2007). The former is likely to be completed online using a browser, whereas the latter is likely to be completed locally and may be uploaded online or may be stored locally (e.g. on a USB Flash drive). In courses such as AIT, which focus on students using computer systems to create artefacts, the latter is likely to be more relevant although not exclusively. There have 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 a few 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 et al., 2007).

The use of computer-based exams to assess IT courses has been used in many places for many years although rarely for high-stakes purposes. More recently, there has been renewed interest with a focus on "21st Century Skills" that typically include assessing capability in the use of computer systems. This trend is clearly seen in the afore-mentioned international research project, the *Assessment and Teaching of 21st Century Skills* project, supported by of Cisco, Intel and Microsoft. In the USA a decision has been made to include an ICT literacy test in national testing in 2012 (Harris, 2008). In Australia a computer-based test was constructed for research purposes to assess the ICT literacy of over 7000 Year 6 and 10 students (MCEETYA., 2007) wherein students were required to use laptop computers to complete a set of tasks within a specially created simulated ICT environment. This was similar to a trial in the UK involving a simulated system to assess the ICT skills of secondary students at the Key Stage 3 level (Boyle, 2006). In both cases it appears that the assessment tasks were successfully implemented; however, they appear to have been too expensive and difficult to manage for widespread use in high-stakes assessment. This sets the stage for the investigation of summative assessment in the Western Australian AIT course.

2. Material and methods

In Western Australia a three-year investigative study was designed to provide adequately authentic assessment tasks for high-stakes purposes in secondary schools – specifically, for performance outcomes that do not lend themselves to pen and paper representation, and in a manner that generates reliable measures, is manageable to implement, and without large increases in cost. One of the components of this study was to assess the performance of students in the AIT course using ICT to develop digital solutions to complex problems.

2.1. Design of study

The study was conducted in three one-year phases: the first phase – proof of concept – reported in this paper; the second phase was a 'prototype' phase, and the third was a 'scale up' phase. The first two phases were concerned with the feasibility of implementing particular forms of performance

assessment with the support of appropriate digital technology, and the third phase was designed to investigate the feasibility of implementing these digital forms of assessment across an educational jurisdiction for the purposes of high-stakes summative assessment. This design was adapted from that of the British e-Scape research project, with its four dimensional framework for analysing feasibility: technical, pedagogic, manageable, and functional dimensions (Kimbell & Wheeler, 2005). The use of an adaptation of this framework led the study's research design to be ethnographic in nature, using interpretive techniques with a combination of qualitative and quantitative data gathered from the main participants: students, teachers and assessors. It was decided to initially take a case study approach, with each class of students as a separate case because previous research had shown that the ethnography of each class was different which affected the feasibility of particular digital forms of assessment (Centre for Schooling and Learning Technologies, 2008). This multi-case approach (Burns, 1996) would then increase the generalisability of findings across types of performance outcomes and forms of assessment.

In 2008 the first phase of the AIT component of the study involved seven teachers, each with one class of senior secondary students, giving a total of 115 students. Firstly a situation analysis was conducted, then assessment tasks were developed, and implemented with each class, and the work output was collected and marked. During and after implementation, data were collected using observations, a survey of the students and teachers involved, interviews of these teachers and students, and interviews with assessors. These data, including achievement data and the results of the marking processes, were analysed for each class and for the sample as a whole. The student questionnaire, teacher questionnaire/interview, and student forum interview proforma were developed for the course based on those developed in the pilot study in 2007 (Centre for Schooling and Learning Technologies, 2008). There was a consistent structure for all data collection instruments across the four courses, with only the content varying depending on the nature of the course and the assessment tasks implemented.

A research team comprising four researchers and three curriculum and assessment support officers from the awarding body, the W.A. Curriculum Council, conducted the situation analysis that was designed to provide a basis for designing an appropriate high quality assessment task. After this the seven teachers were recruited and added to the team to assist in refining the assessment task ready for implementing. Teachers were recruited on the basis that they were experienced in teaching the course and would agree to implement the assessment tasks for one of their classes within a program that would accommodate the tasks.

2.2. Developing the assessment tasks

The AIT research team was responsible for developing the assessment tasks. Their aim was to develop authentic performance assessment tasks that met the required standards of the Curriculum Council and could be readily implemented in a school situation. Initially a situation analysis was conducted to consider what was possible within the requirements of the course, the performance requirements, the potential technologies, the constraints of the school environment, and teacher and student characteristics. However, the aim was to test the boundaries by considering the use of technologies as close to the 'cutting edge' as possible. As a result of this process, the team identified the content and outcomes conducive to digital forms of assessment for the course and drafted appropriate assessment tasks. Finally, marking criteria and marking keys were developed and these, along with the assessment tasks, were reviewed by a group of curriculum and assessment experts.

The research team decided to implement two main forms of assessment in the first year: a digital portfolio and a computer-based performance exam; hereafter referred to as the *Portfolio* and the *Exam*. It was considered that both forms would meet the requirements identified in the situation analysis and there was a need to compare the feasibility of each. These assessment tasks were defined in terms of five components, three for the portfolio and two for the performance exam, as indicated below.

Portfolio – a digital portfolio constructed during 20 hours over 5 weeks

> *1: Digital Product – a prototype of an information solution using applications commonly used in organizations for productivity, planning and communication*

2: Process Document – a document related to the digital product, collated over a period of five hours with a maximum of nine pages that comprised four sections: research, design, production and evaluation

3: Extra Artefacts – previously created at school under supervision, that illustrated skills in applying design principles in any two domains (e.g. graphics, databases, spreadsheets, web-publishing)

Exam – a three-hour computer-based exam

4: Reflective Questions – a set of reflective questions concerning the portfolio digital product

5: Performance Tasks – a set of six tasks provided as a scaffold for responding to a design brief – the tasks involved producing a brochure for a holiday resort, including creating a logo, graphs and tables

For each component a set of marking criteria was developed as a rubric-style analytical marking key. These criteria were based on the tasks and the requirements of the course syllabus. Teachers were not required to use this marking key but were permitted to do so for their school-based assessment. Later, three criteria were distilled from the analytical criteria for use with the comparative-pairs marking.

2.3. Task implementation

The digital portfolio and computer-based performance exam were implemented in the seven classes during the second half of 2008. The results of the analysis of data related to this implementation are now discussed, starting with a presentation of the method of implementation of the task and the technologies employed. This is followed by results of an analysis of the data collected from marking the students' work and from surveys and/or interviews of students, teachers and assessors. Results for each school are not discussed separately here but are presented as case studies in the first official report for the partner organisation, the Curriculum Council of WA (Centre for Schooling and Learning Technologies, 2009).

Each class was visited at least four times: typically at the beginning of the portfolio, towards the end of the production component or beginning of the process document component, during the implementation of the exam and after the exam. All of the sessions when students worked on the assessment task were held in a computer laboratory at the school, and facilitated by the teacher. Students in all seven classes attempted both the portfolio and exam; however, for one class the portfolio was not submitted and in most of the other classes the extent to which individual students completed all components of the portfolio varied considerably. The exam was completed with almost no technical difficulties evident, apart from the recording of sound (used for students to present their reflections on their designed prototype) for three of the seven classes. For each class the teacher facilitated the portfolio development, and a researcher and the teacher administered the exam.

2.3.1. Portfolio

The design brief for the portfolio *Digital Product* allowed teachers to insert a scenario, including type of product; however, four used the example provided in the project support documentation, *The Miss Shoppe* website. The focus of the activity was the application of the whole technology process to a real-world context, as set out in the scenario contained in the design brief. Students had 15 hours of class time over four weeks to develop a prototype product. Students were required to complete all work during class time but some teachers did not adequately invigilate this with clearly some students completing some work at home. Hardware and software were restricted to those available at the school.

On completion of the *Digital Product* students were supposed to collate evidence of the investigation, design, production and evaluation processes undertaken into a *Design Process Document* for which students had five hours of class time. Students also submitted two additional *Digital Artefacts* they had created in the course, along with two half-page forms explaining the artefacts. It was intended that these should demonstrate ICT skills and knowledge over and above that represented in the *Digital Product*, but this was generally not emphasised by teachers.

Typically student portfolio work was delivered to the researcher by the teacher on a disc and organised with a folder for each student. The enclosed files were transferred, by a researcher, to the student

folders on the project server with files named in a consistent fashion (i.e. the same file name for the same purpose for each student).

2.3.2. Exam

Typically students completed both components of the *Exam* contiguously over three hours under 'examination' conditions, with the teacher and researcher invigilating. Students were given a paper copy of the examination, a 4GB USB flash drive and an audio headset with microphone. The teacher was responsible for setting up the workstations, and the researcher provided everything else. There was 10 minutes reading time prior to the commencement. Students from the first class to complete the exam were required to do the reflective questions first for an hour and then the performance tasks for two hours. However, it appeared that students were not happy with this arrangement, wanting to move to the performance tasks more quickly. Therefore, for all the other classes the two sections were reversed with the two-hour performance tasks preceding the reflective questions. Students were not permitted to continue with the performance tasks once the allocated two hours had expired.

Students were asked to type their answers to the *Reflective Questions* component of the exam into a *Microsoft Word* document provided on the USB flash drive. The questions asked them to reflect on the *Digital Product* development component of the portfolio. For the *Performance Tasks* component of the exam, students were given a real-world design brief and prompted to follow a technology process to create a digital product.

With the exception of design sketches, which had the option of being paper or computer based, the entire examination was done on computer; students' responses were saved as digital files in various formats. The USB flash drive contained 18 digital photographs, a text file of data, design templates as .doc and .ppt files and a .doc template for preparation of an audio reflection. An A3 size printed copy of the design template was also supplied to give students the option of designing on paper. Student design work that was done on paper was collected and either scanned or photographed to add to their digital work. Students were permitted to use any software available on their desktop computer and to save their work to the USB flash drive – typically a copy was also saved to the school's server. Most students used common office and graphic production software such as those provided by Microsoft and Adobe. Later a researcher transferred all digital work from the USB flash drives to a project server as the online repository.

2.3.3. Online repository

The assessment outputs for all students were uploaded to the online repository so that the work could be accessed online by markers. In the AIT repository all files stored on a project server in a unique folder for each student, named using the student's ID. Each student's folder contained a folder for each of the artefacts, one for the portfolio product and one for the exam. The portfolio process document was a PDF file, which was placed within the student's main folder on its own. Within each folder there was an index.htm page that was used by the marking tools to display the contents of the folder, with links to the other files. This folders and files structure was set up manually. The first artefact folder contained a PDF file of the student's descriptions of the two artefacts. The exam folder contained all the files copied from the exam USB flash drives and a PDF file combining the design plans, brochure, and reflections.

2.3.4. Marking tools

Two marking tools were developed using the FileMaker Pro relational database software to facilitate the analytical and comparative-pairs marking. The marking criteria were part of these online marking tools. The tools were designed to display student work output and facility to input assessor judgements, all within a 20" screen. The FileMaker Pro software allowed the tools to be deployed on the Internet with minor modifications, and to provide unique password protected logons for each assessor.

The analytical marking tool had the assessment criteria displayed on the left side of the screen and the student work on the right. The tool consisted of a *Student Results List* screen and five 'marking' screens, each with different criteria and the appropriate student work displayed. Marks were recorded by clicking on buttons, as was navigation between marking windows, and assessor notes could be typed into a field when required. The tool was also designed to do all the clerical functions, such as totalling the marks.

The comparative-pairs marking tool was designed to display two students' work side-by-side on the screen, with the recording of the assessor's choices located between them. The assessor was required to make four choices, associated with one holistic and three specific criteria, by clicking on large green arrows pointing to the student work they judged to be superior. A short description was given of what to consider for each of the judgements. A field was added for each student to allow assessors to type notes on the students' work so that they would not have to fully review a student's work after the first occurrence. When completed, assessors clicked on a button to go to the next pair to judge. From the assessors perspective, the comparative-pairs marking tool consisted of a *Student Results List* screen and a *Which Is Best* screen. Pairs of student work to be judged were preloaded for each assessor. These pairs had been determined using a standard statistical randomisation procedure for the comparative pairs method.

Two external assessors were recruited to complete the analytical marking; these two plus another three assessors conducted the comparative pairs marking. All assessors were experienced computing teachers with considerable experience with the course.

3. Results and discussion

Data were analysed for each case study (class) and then for the combined group as a cohort. The full range of research data was sought for each case study; however, for each case there were data missing. For example, no teacher provided a full set of marks although all but one provided a set of semester marks. The survey and interviews were conducted for all case studies. For one case study no portfolios were submitted. In the final analyses these omissions were treated as missing data.

3.1 Observations, surveys, and interviews

The students in each class were observed a number of times while they were completing the portfolio and exam. Anecdotal records were kept of their behaviours by a researcher and collated into a table for each class. These data were then analysed across the cases by looking for consistencies and variations. A photograph was taken of the computer laboratory being used. The survey of students and the student forum interviews were conducted as soon as possible after the completion of the exam. The forum involved at least one small group of students for each class. These group interviews were audio recorded and then notes made to identify key points. The assessors were informally interviewed with notes made to identify key points. The teachers were 'interviewed' using emailed questions after marking was completed and they were also asked for feedback after they had received their case study report. Responses were collated by question and then common themes extracted. It is not practicable here to report all the results from these analyses; results critical to this paper are reported here and in summary Table 5.

Five of the seven teachers ran the portfolio in whole or in part as an additional task, not counting towards a students' final semester mark. Being a research project teachers could not be required to include any of the assessment tasks within their grading scheme. It was clear from observations and responses to interview questions that many students did not give the portfolio their best efforts and as a result many portfolio submissions were incomplete. Only 44% of students submitted all components of the portfolio, with the best being 71% submitting the *Digital Product* component. Most of those not submitting the *Digital Product* component were from one class where the students completed a product but for reasons that are not entirely clear the teacher was not able to deliver the resulting files for marking. The computer-based exam was used within the school-based assessment by six of the teachers, with the other promoting it positively with her students. This resulted in most students appearing to take the exam seriously, with a higher standard of work on average, as noted by assessors.

The survey of students aimed to capture student perceptions on the effectiveness of the exam and portfolio, their use of computers and other digital devices at home and school, their attitudes to using computers, and their capability of using a range of computer applications. The questionnaire consisted of 58 closed-response items and two open-response items. It was based on the questionnaire used by the British e-Scape project and questionnaires used in previous projects at the Centre, and was trialled in the pilot project (Centre for Schooling and Learning Technologies, 2008). In total 110 students completed the questionnaire, with responses collated for each case study and for the entire cohort.

Basic descriptive statistics were calculated for each closed response item and responses to the open-response items were transcribed into tables for each case, summarised into a reduced set of responses and then compared across cases to determine consistencies and variations. Seven scales were derived from combining closed-response items. Basic descriptive statistics and frequency histograms were generated for each scale (see Table 1 and Figure 1). These scale scores were later used to identify groups of students and to test for relationships with scores from marking.

** TABLE 1 **

** FIGURE 1 **

On the questionnaire students were asked two questions about the amount of experience they had had with computer-based exams and digital portfolios and how quickly they could adapt to them. Most students indicated little experience with computer-based exams (44% no experience) and a substantial proportion with digital portfolios (17% no experience and 25% little). However, the majority (62% exams and 51% portfolios) felt they would need little time to get used to these approaches. Only 6-8% of the students indicated that it would take 'lots' of time to get used to these approaches. The eAssess and eAssessP scales each combined responses to 11 items on a 4-point Likert response pattern. The scales purported to measure the ease of completion and efficacy of each form of assessment. Both scales had high Cronbach-Alpha reliabilities (Table 1) and means of 3.2 and standard deviations of 0.4. These means were well above the mid-point of 2.5, with frequency distributions skewed positively (Figure 1). In general almost all students indicated a positive perception of both the exam and portfolio and a preference for these to pen and paper examinations. They considered these forms of assessment to be quick, easy and good for representing their capabilities in the course. For example, 92% indicated that "the computer is a good tool for designing products in an exam" and 94% for the portfolio. Clearly the experience of completing the portfolio and exam had been positive for these students.

The questionnaire also had items concerning access to technologies at home, their use of digital technologies at school, and their experience, attitudes and skills in the use of particular applications of these technologies. Nearly all the students had home access to the technologies listed in the questionnaire, with mobile phone and MP3 player ownership both more than 90%. Two thirds of the students owned their own laptop computer, 95% had a broadband Internet connection, and 87% indicated using a computer at home on a daily basis. They estimated that on average computers were used 95 minutes per day at school, which is not surprising for an IT course (this compares with about 40 minutes for the other three courses in the main study). Although the *Apply* and *Confidence* scales were not very reliable (Table 1), consideration of responses to individual items supported the contention that generally these students indicated being positive and confident about using computers. For example, 75% indicated that they enjoyed using computers at school and 88% indicated that they felt confident at using computers. More than 80% of students indicated competence on the 11 types of software applications in the skills rubric, except for spreadsheets (79%), web authoring (75%), digital video (72%) and databases (60%). The *Skills* scale constructed from these items gave a mean of 3.3 on a four-point scale (Table 1). This compared very favourably with the means for the other three courses in the main study (all between 2.9 and 3.1).

It was deemed important to determine the extent to which the intentions of the assessment tasks had been understood by students and teachers, with evidence derived from observation, marking and comments made by students and teachers. The portfolio was familiar to all teachers and most students, it being very similar to what would typically be done as classroom exercises in the course. The performance tasks test was less familiar but appeared to be clearly understood by teachers and students. The ease with which almost all students created a business logo and a tri-fold advertising brochure illustrated this familiarity. However, the creation of graphs using a spreadsheet appeared to be less familiar to students from some of the classes (e.g. overall 36% of students received the highest judgement for the spreadsheet task, with only two classes having below that proportion) of students, probably indicating little recent experience. Further, there was widespread variation in students' interpretation of what constituted a logo with more than half the students simply adding a caption to one of the photographs supplied. Fewer than 20 students designed a logo and used drawing tools to

create it, and even here many used sections from the photographs supplied. The pervasive use of built-in templates and wizards in software such as Microsoft Publisher was further indication that the main task was familiar for students. Only five students didn't use a template, with four of these using a three-column word-processed document. In general, students selected appropriate photographs for the brochure showing that they understood the intention of the task was to market a resort as luxurious but having a low environmental impact. Student reflections supported this contention.

The reflective questions component of the exam was not well done, indicating that the intention of the questions was unclear for students. One student commented "The wording in the second part…a bit confusing. I had to guess at what it meant", and another, "It took me a while to work out what was meant by it". There was widespread confusion over the stages of the technology process and the distinction between these, with many responses repeated. A student explained that, "it just seemed like you asked the same questions four times…I got four words out of a thesaurus and copied and pasted those in three or four times".

In general the teachers were very positive about the assessment by portfolio and by practical examination and felt that these complemented their own aims, principles and methods of instruction. As one teacher observed, "If the external marking of a portfolio does away with the moderation process, I'm all for it". All teachers said they would like to see a greater emphasis on the practical aspects of the course. One teacher commented, "If we are asking our students to complete the majority of their assessments using these tools throughout the year, then surely we should in the final exam". Some teachers were cautious about the potential of the examination because of the possibility of technical problems. One teacher suggested running the examination from a bootable mass storage memory device, such as a USB flash drive, containing not only the data files but also the application software, and commented, "In this way there is more control over the whole environment".

3.2 Marking process

The two external assessors who completed the analytical marking and the three others who joined for the comparative pairs marking were interviewed to gain their impressions of the marking process, quality of student work, time taken to mark the work, and operation of the digital repository and online marking tools. The marking process was simplified by the fact that all submissions were in digital form, allowing anytime anywhere access, and by the use of the online marking tool. Assessors generally appreciated these features but commented on some limitations with the marking system such as delays in opening large files and scrolling between the mark key and work sample. Changing a mark already entered was a little clumsy requiring a post back of the marking form. The running score of the mark also didn't update until the marking form was submitted and this was confusing at first. The assessors believed that the quality of student work varied widely and this supported the suitability of the tasks as discriminators of student ability. One assessor alluded to the common misunderstanding, in the examination, of what was meant by a logo, and suggested that some examples would have aided clarification.

The amount of time taken for the analytical marking ranged from about 5 minutes to 25 minutes per student for the portfolio and exam. The time taken tended to depend on the completeness of the student's work, the quality of the assessor's access to the Internet, and the size, type and complexity of the files that needed to be accessed. For example, some animations and videos were more than 5MB. Higher quality work often took longer to mark, with evidence of performance being sought from many parts of the work and greater consideration required in making judgements.

The comparative pairs marking focussed only on the performance tasks exam component, so the time per student was correspondingly reduced. The time required to make a comparison was initially around 10 minutes, mainly where the samples were of similar quality. However, as familiarity with the criteria increased the time per pair became less. Because the comparisons were pre-determined and not dynamically generated, several were very one sided and for these the marking time was seconds rather than minutes. For the comparative pairs marking assessors took on average about 3 minutes per comparison and made 354 comparisons involving the work of 60 students, resulting in an average time of 18 minutes per student.

3.3 Results of marking

The two external assessors marked all of work submitted for the 115 students. The two marks for each component were averaged between the assessors and then totalled. The comparative pairs marking only included the performance tasks exam component for a reduced sample of 60 students. These students were chosen because their practical work samples were equivalent in the degree of completeness and had no missing sections. In particular they all had an audio response file for the exam. Five assessors each completed the same pre-determined set of comparisons between students using a digital marking tool. Of the 115 students, only 58 final semester and 26 assessment task marks were received from teachers.

For each case study the results of marking were compiled into a table showing the scores for each individual student from analytical marking, comparative pairs marking and teacher marking. For each method of marking each student was also given a ranking. The ranking for the average of the analytical marking was based on all 115 students, whereas the teacher's semester mark rank was just within the class. The ranking from the comparative pairs marking was for the 60 students whose exam was marked in this manner. Correlations between these scores and rankings are summarised in Table 2.

** TABLE 2 **

3.3.1 Results from analytical marking

There was a strong and significant correlation between the scores of the two assessors for the overall cohort, with correlation coefficients of 0.89 ($p<0.01$) for the scores and 0.91 ($p<0.01$) for the ranking of *All* students (Table 2). This was also the case for five of the classes. Not only were the scores generated by the two assessors highly correlated, there was no significant difference (t test) between their means on the total mark (Portfolio and Exam combined) and they generated very similar ranges of scores and standard deviations. However, the range of scores, the means and standard deviations did vary considerably between classes on separate components and for the assessment task as a whole. For example, Class ZA had a mean of 53.6 while Class RA had a mean of 31.2.

When compared with scores submitted by teachers who used their own analytical marking schemes, there were no significant correlation between the analytical marking scores and the scores awarded by the teachers ($r=0.32$), nor for the ranking (Table 2). However, when compared with semester marks awarded by teachers there was a moderately strong and significant ($r=0.62$, $p<0.01$) correlation. This would indicate that in general students' scores on the assessment tasks were in line with achievement throughout the course but that on a particular assessment the nature of the marking criteria used was critical. That is, overall external scores for the portfolio and exam reflected student achievement throughout the course but because the teachers did not use the provided marking rubric their scoring was not consistent with that of the external assessors.

The assessment task had two major components, the portfolio and the exam, each with sub-components. The results of marking of each component were analysed separately, with some of the summary statistics for the portfolio and exam shown in Table 3 for the external 'analytical' assessors and the teachers. The mean score for the external assessors for the exam was around 50% but for the portfolio only around 37%. This discrepancy is probably indicative of the lack of effort of students on the portfolio. Of the exam work marked by the teachers, the mean was quite similar at 52.5%. There was a moderate but significant correlation between scores awarded by the external assessors for the exam and for the portfolio ($r=0.58$, $p<0.01$) but very weak ($r= 0.36$, $p<0.01$) when compared with marks awarded by the teachers. The analysis on rankings delivered similar results. These results once again tend to indicate that the external assessors were able to consistently apply the marking criteria for all components of the assessment, and that students achieved similar results relative to each other for the portfolio and exam.

** TABLE 3 **

An ANOVA (Analysis of Variance) was conducted to consider variance on the results between the classes of students. With each class being taught by a different teacher it was expected that results of marking would vary considerably for the portfolio and exam components. The analysis indicated

significant variation ($p<0.01$) by class for the portfolio and exam and components of each. These variations were generally consistent with those classes, with higher means for the portfolio also having higher means for the exam. This would indicate underlying differences in the capability of students between classes. However, there were differences between the classes in the extent to which the results of analytical marking were correlated between the exam and portfolio. For example, the ZA class marks were more highly correlated ($r=0.55$, $p<0.05$) than for XA ($r=0.31$, $p>0.05$), probably explained by the more stringent implementation of the portfolio requirements for the former class.

3.3.2 Results from comparative pairs marking

For the comparative pairs method of marking one holistic and three specific assessment criteria were developed based on the analytical marking criteria for the performance tasks component of the exam.

Holistic criterion	Brochure is effective for target customers through developed planning to incorporate all the required features and information, appropriate use of aesthetic effects on a theme, consistent and balanced layout, and professional look.
Specific criterion 1	*Design Process*: Product originates from planned design showing development of ideas and justification in reflection.
Specific criterion 2	*Technical Proficiency*: Demonstrable capability and facility with the range of required software (spreadsheet, logo, brochure).
Specific criterion 3	*Design Principles*: Creative application of appropriate design principles and elements such as alignment, balance, contrast, emphasis, harmony, proportion, proximity, repetition, unity, and white space.

The marking involved five assessors, each making judgements on 354 pairs of student work, using an online marking tool. The four sets of scores (i.e. based on judgements for the four criteria) were exported into spreadsheets and subsequently imported into the RUMMcc (Andrich, Sheridan, & Luo, 2003) software specifically designed to analyse data from comparative pairs marking. This software was used to apply a dichotomous Rasch model to generate a single score for each student in logits (logarithmic units of measurement) with a standard error of measurement. A Separation Index (SI value between 0 and 1) was calculated as an indicator as to whether or not the exemplars (student work) were sufficiently diverse in quality to assure a broad enough range for the purposes of comparison. The SI for the holistic criterion was 0.96, indicating a highly reliable set of scores (values above 0.8 are considered to be good). Intra-rater reliability analysis gave a group reliability was 1.01 where this statistic should be between 0.5 and 1.5. Results for the three specific criteria were similar to those for the holistic criterion.

Correlation analysis was conducted to investigate the relationship between the scores generated for the four criteria and between these scores and those generated by the analytical 'external' marking and the teachers. Strong and significant correlations were found between each of the three specific criteria and the holistic criteria and between the three themselves, with the weakest correlation being between criterion 1 and 2 ($r=0.74$, $p<0.01$). Similarly there was a strong and significant correlation ($r=0.73$ $p<0.01$) between the scores generated by comparative pairs marking and those generated by analytical marking. However, there was no significant correlation between the scores generated by the teachers for the exam and those generated by the external assessors using either method of marking, with the exception of criterion 2 in the comparative pairs marking ($r=0.461$ $p<0.05$). This indicates that the teachers tended to assess the exam more in terms of technical proficiency. There were weak correlations between the external scores and semester scores provided by teachers.

A similar analysis was performed using the rankings produced by each of the marking methods. The rank on analytic marking was strongly and significantly correlated with all criteria of the comparative pairs marking. The strongest correlation was between the analytical score and the holistic criterion score ($r=0.71$, $p<0.01$). There was again no significant correlation between the rank of teacher's examination score and the comparative pairs scores, with the exception of criterion 1 ($r=0.43$, $p<0.05$). The teacher's semester mark was weakly correlated with scores on all criteria in the comparative pairs marking, with the holistic criterion being strongest ($r=0.53$, $p<0.01$).

3.3.3 Applying a polytomous Rasch model to the results of analytical marking

A polytomous Rasch model was applied to the *Exam* and *Portfolio* analytical marking using the judgements of both assessors for each criterion. The data were analysed using the RUMM2020 (Andrich, Sheridan, & Luo, 2006) software package designed for applying a range of Rasch models. This analysis identified one reversed threshold for one *Exam* criteria and as a result the analysis was repeated with two responses to that criterion combined. This increased the SI marginally to 0.85 (Cronbach Alpha 0.85) and removed the reversed threshold. For the two components of the exam there were few extreme outliers, with the frequency distribution relatively well spread (see Figure 2 for distribution for the *Performance Tasks* component of the exam). The analysis gave a reliable set of scores for all three components of the portfolio (SI=0.96, 0.96 and 0.92 respectively, with Cronbach Alpha coefficients of 0.94, 0.96 and 0.94 respectively). There were a few extreme outliers, particularly for the first component, the *Digital Product*. These tended to be students scoring 0 on all or almost all of the criteria. The frequency distributions tended to be well spread, with high standard deviations and not very 'normal' in structure. The thresholds on all items worked adequately. No modifications were required, although the thresholds for three of the criteria did not discriminate well. In general the analytical marking of the exam and portfolio components generated reliable scores. However, some improvements in some of the marking criteria could be made. The SI for the performance tasks exam component was significantly lower than for the three components of the portfolio, suggesting that the latter were slightly more reliable measures.

** FIGURE 2 **

3.3.4 Reliability of Exam scores compared with Portfolio scores

This section considers in more depth a comparison between the reliability of the scores generated by the marking of the *Portfolio* (all three components combined) and the *Exam,* using inter-rater correlation as a measure of reliability. Only the *Performance Tasks* component of the *Exam* was marked using both methods of marking so this sample of 60 students is considered. Table 4 shows correlations for this component and the *Portfolio*, a similar analysis of rankings rather than scores that gave similar results. There was a relatively moderate but significant correlation ($r=0.43$, $p<0.01$) between the two external markers on the *Exam* scores but a high correlation ($r=0.93$, $p<0.01$) for the *Portfolio*. However, their average scores (*Exam Analytical* in Table 4) were relatively highly correlated to the results of the comparative pairs marking for the Holistic criterion. In general there were only moderate to low significant correlations between the scores for the *Exam* and the *Portfolio*. These results would seem to support the conclusion that the scores generated by the analytical marking of the *Exam* were significantly less reliable than for the *Portfolio*. As noted earlier, the comparative pairs method of marking the *Exam* generated a highly reliable set of scores.

** TABLE 4 **

Only two schools (MA and ZA) implemented all aspects of the portfolio in line with the stated requirements; therefore, this sample of students was analysed separately. Figure 3 shows a graph of the spread of scores for these two classes combined, with the *Exam* and *Portfolio* scores significantly correlated ($r=0.71$, $p<0.01$). The correlation between the two external assessors using the analytical method was significant at the 0.01 level for the *Portfolio* ($r=0.85$) and *Exam* ($r=0.54$). For this sample this measure of reliability was still much more acceptable for the *Portfolio* than for the *Exam*.

** FIGURE 3 **

To investigate the seemingly lower reliability of the marking of the *Exam* compared with the *Portfolio,* an analysis of the operation of the analytical marking criteria for the *Performance Tasks* component of the *Exam* was conducted. Using t tests to test for differences in means between the assessors, a significant difference was found for the overall score but only a significant difference ($p <$

0.05) on two out of the seven separate analytical criteria. For one of these criteria assessors were to allocate 0, 1 or 2 in response to "Applies appropriate file formats, compression and encryption techniques, conversion, size and storage requirements." Generally no scores of 2 were allocated, just scores of 1 or 0. The second criterion was associated with the production of the brochure, "Used specific styles, forms and techniques to create brochure represent the particular effect of the design on the audience and achieve defined standards of quality." Once again the maximum possible score of 4 was sparingly used by both assessors. It was concluded that in the future an emphasis should be placed on defining the highest level of achievement for a criterion to encourage assessors to select it.

3.4 Feasibility analysis of Portfolio and Exam

A feasibility framework based on the work of Kimbell, Wheeler, Miller and Pollitt (2007) was used to compile a summary of the findings for the *Portfolio* and the *Exam*. This was distilled from the results of an analysis of the range of data across all the cases. A summary of the results of this exercise is shown in Table 5.

** TABLE 5 **

In general from an analysis of the 2008 data associated with the AIT course it could be concluded that both the *Portfolio* and *Exam* were appropriate measures for assessing student achievement in the course. Almost all participants considered both forms to be considerably more authentic than a paper-based exam. Both assessments were feasible to implement in schools but the *Exam* could be more consistently implemented because it was short, narrowly focussed and did not rely on teacher invigilation. However, using the analytical method of marking the *Portfolio* could be more reliably marked, with the *Exam* requiring the comparative pairs method of marking to provide a comparably reliable set of scores. Rasch analysis indicated that the *Exam* scores were highly reliable (SI=0.93) using the comparative pairs method of marking and, with a minor modification to one criterion reasonably reliable (SI=0.85) using the analytical method of marking, but less reliable than for the *Portfolio* scores (SI=0.96).

There were a few manageability and technical issues connected with the implementation of the assessment tasks and marking of student work output. For the *Portfolio* there were no serious technical issues but the inconsistent implementation by teachers gave rise to both manageability and functional concerns. Although a set of parameters and guidelines were provided, only two of the teachers appeared to largely comply. For example, most did not allocate the correct amount of time, usually allowing more time; and most did not include all components within their assessment schedule and some allowed students to complete work at home. Many students did not submit all required files. These outcomes seriously reduced the validity of the portfolio as a measure of achievement. For the *Exam* the only technical issues concerned the audio recording and a few malfunctioning workstations. For various reasons students in three classes were not able to complete the audio recording and a few students had to move workstations during the exam time due to technical failures, resulting in the loss of no more than five minutes of work for each.

There were difficulties in preparing student work from the *Portfolio* and *Exam* for marking, with some file conversion and renaming required. Most often students did not provide the files in the format or use the name specified in the instructions. Often this appeared to be because they did not know how to do so, but sometimes it appeared this was due to lack of diligence or memory. In some situations submitted files were converted into alternative formats for ease of access through the marking tools. For example, where files could be displayed as PDF, HTM or MOV files (e.g. documents, slideshows and spreadsheets) they were added to the repository to complement the original file. Typically the assessors viewed the reformatted files because they could be displayed through the browser, but, when necessary, they downloaded original files. This ease of access made the effort in reformatting and renaming worthwhile and would be increasingly so with larger numbers of students and the use of batch processing of files.

4. Conclusions

This paper has reported on a component of the first phase of a study concerned with improving the pedagogical alignment and authenticity of the high-stakes summative assessment for a senior secondary school applied IT course. In this phase a digital portfolio comprising three major components and a computer-based examination comprising a one-hour short typed responses component and a two-hour practical performance tasks component, were implemented with seven classes of students. All student work was collated in digital form, which facilitated ease of storage, transmission and access. All student work was judged by external assessors using browser-based tools accessing an online repository to implement an analytical marking method and a comparative pairs marking method.

The computer-based examination was implemented for all students in the seven classes, with only minor feasibility concerns apart from the recording of audio reflections for three classes. However, for the digital portfolio only a minority of students submitted all components of the portfolio and in general completion varied considerably. This was most likely due to most teachers not including the portfolio within their official assessment schedule, whereas all but one did so for the exam. Furthermore, it was clear that the teachers had difficulty invigilating the portfolio but the exam was externally invigilated. The resulting digital files from both forms of assessment needed a reasonable amount of checking, reformatting and renaming to ensure consistency for uploading to the online repository.

The online database marking tools, which were used to access the student work by assessors, were relatively easy to use and responsive if an assessor had reasonably fast Internet access, was not unreasonably limited by firewalls, and had workstations with adequate processing speed, memory and screen size. Marking was possible, and indeed took place, from countries outside Australia though opening of large files presented delays. Markers within the Curriculum Council's network experienced difficulties with accessing some files, marking tools operating slowly, and sometimes software crashing or system log outs. With regard to the analytic method of marking, the ability to view both the work output with the marking rubric alongside was convenient for assessors switching rapidly between different aspects of student work. The database recorded and collated the scores from such judgements and allowed these to be quickly and accurately extracted for analysis. The analysis of the scores from analytical marking, with the strong correlations between assessors, indicated that the method had generated reasonably reliable scores. After some practice with the system and a brief familiarisation with the criteria, the comparative pairs marking tool was also quick and convenient. Rasch analysis of assessor judgements indicated that both methods of marking generated scores with more than adequate levels of reliability. However, more in-depth analysis found that the analytical marking of the portfolio was more reliable than for the exam.

When comparing the overall feasibility of the digital portfolio and the computer-based exam using the feasibility framework, there was no compelling reason to choose one over the other. The portfolio assessment best aligned with the intended pedagogy of the course, could be more reliably marked, potentially had higher content and construct validity if implemented rigorously, and was more flexible to allow choice of context for students. On the other hand, the exam was more manageable and was easier to standardise to support criterion-related validity. Due to the short working time available in the exam, it was not able to assess the relevant intended outcomes of the course as comprehensively or validly as for the portfolio. Although scores generated by analytical marking of the exam were not as reliable as those for the portfolio when a comparative pairs method of marking was used, the scores generated were highly reliable. Thus overall it could be concluded that students were better able to demonstrate their capability through the portfolio but this was less manageable for high-stakes assessment. It could be argued from these results that both forms of assessment had greater authenticity than a purely paper-based examination with the portfolio having better alignment with the curriculum and pedagogical practices in the course.

To address the shortcomings of the portfolio an online portfolio management system would be needed to support a well-structured and tightly controlled system for consistency and verification. In addition some type of signed affidavit with spot checks on a sample of students would be needed to ensure all teachers implemented the portfolio according to the required conditions.

To address the shortcomings of the computer-based exam there are two areas in which further consideration is required: firstly, the technical requirements for a full-scale trial; and secondly, improving the content and construct validity of the assessment task or the specifications of the performance tasks. For the study the exam was implemented using USB flash drives on standalone school-based computer workstations. Although it would seem easier to implement using online technologies it was determined that the reliance on school networks was too risky at this stage. Although USB flash drives are a little cumbersome, it is not unrealistic to countenance their mass distribution and collection of USB flash drives. Another manageability, technical and functional consideration is whether limitations are placed on the software students may use. The study allowed students to use any software normally available on their workstations. With the focus on design and relatively low-level IT skills, there was no evidence that any of the students were disadvantaged by software availability. Most students used standard Microsoft or Adobe software. Further, the marking criteria did not consider skills levels but rather the application of techniques and skills to the requirements of the design.

Compared with the portfolio, the exam provided much less scope for students to demonstrate their capabilities in using the technology to design and develop a product. In the exam students were only required to create a graphic logo, a spreadsheet graph and a tri-fold brochure, all relatively low-skill tasks for upper secondary students. These were chosen because it was considered that almost all students would be able to attempt these using a typically standard set of software that they would all have available. The situation analysis conducted at the beginning of the study concluded that there was a vast range of types of practical tasks that teachers gave their students, and therefore it was not possible to set more difficult tasks that students would have had the background experience to tackle. This was not a problem with the portfolio as the design brief could vary between classes and allowed tasks to be relatively open-ended. It was therefore decided that for the second year of the study the students in the computer-based exam would be given a choice of two tasks to allow greater complexity.

The findings reported in this paper were used to inform decisions made for the second year of the study in which the assessment tasks were refined and the implementation processes improved. For example, for the AIT assessment task this resulted in the removal of the reflective questions component of the exam and the audio recording of reflections in the performance tasks component of the exam. Furthermore, in this component students were less constrained by being given some choice of product type. The portfolio structure and requirements remained the same, but there was an increased focus on encouraging teachers to include it within their formal assessment schedules and to more rigorously invigilate compliance with requirements. It was also decided to use an online portfolio management system to support students in submitting their work. The intention then in the third year was to trial the assessment tasks with a broader range of types of schools to demonstrate the feasibility for state-wide implementation in the future. Thus the aim throughout the three years was to improve the authenticity of the summative assessment, to reward appropriate pedagogy, and to support the delivery of a high quality, more relevant course for students in Western Australia. The first year of this study demonstrated the strengths and limitations of two digital forms of assessment for the course. These findings will be of relevance to wider jurisdictions and a large range of other types of courses. For the authenticity of summative assessment needs to be of concern in all education systems and it is likely that increasingly digital forms of assessment will provide vehicles for much needed improvement.

Acknowledgement

<separated>

References

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

Andrich, D., Sheridan, B., & Luo, G. (2003). RUMMcc. Perth, Australia: RUMM Laboratory.

Andrich, D., Sheridan, B., & Luo, G. (2006). RUMM2020. Perth, Australia: RUMM Laboratory.

Barrett, H. C. (2007). Researching Electronic Portfolios and Learner Engagement. *Journal of Adolescent & Adult Literacy, 50*(6), 436-449.
BBC. (2009). Norway tests laptop exam scheme. *BBC News*. Retrieved from http://news.bbc.co.uk/2/hi/technology/8027300.stm
Becta. (2006). *Becta's View: E-assessment and e-portfolios*. Coventry: Becta ICT Research.
Beetham, H. (2005). *e-portfolios in post-16 learning in the UK: developments, issues and opportunities.* Bristol, UK: Joint Information Systems Committee (JISC).
Boyle, A. (2006). *An evaluation of the decision to base the key stage 3 ICT test on a bespoke virtual desktop environment.* London, UK.: Qualifications and Curriculum Authority.
Brewer, C. A. (2004). Near Real-Time Assessment of Student Learning and Understanding in Biology Courses. *Bioscience, 54*(11), 1034.
Burns, R. B. (1996). *Introduction to research methods.* South Melbourne, Australia: Addison Wesley Longman Australia Pty. Limited.
Carbol, B. (2007). *Transition to Online Testing: An ROI Analysis.* Kelowna, BC: Society for the Advancement of Excellence in Education.
Carney, J. (2004, April 14). *Setting an agenda for electronic portfolio research: a framework for evaluating portfolio literature.* Paper presented at the American Educational Research Association Conference.
Centre for Schooling and Learning Technologies. (2008). *Digitally based formats for alternative external assessment for senior secondary school courses in W.A.* Perth: Edith Cowan University.
Centre for Schooling and Learning Technologies. (2009). *Investigating the feasibility of using digital representations of work for authentic and reliable performance assessment in senior secondary school courses. First Year Report.* Perth: Edith Cowan University.
Cisco, Intel, & Microsoft. (2009). Transforming Education: Assessing and Teaching 21st Century Skills. Retrieved 8/6/2009, 2009
Clarke-Midura, J., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research on Technology in Education, 42*(3), 309-328.
Dede, C. (2003). No cliche left behind: why education policy is not like the movies. *Educational Technology, 43*(2), 5-10.
Garmire, E., & Pearson, G. (Eds.). (2006). *Tech Tally: Approaches to Assessing Technological Literacy.* Washington: National Academy Press.
Harding, R. (2006). What have examinations got to do with computers in education? *Journal of Computer Assisted Learning, 17*(3), 322–328.
Harris, S. (2008). Governing Board Awards WestEd $1.86 Million Contract To Develop First-Ever Technological Literacy Framework. Retrieved from http://www.nagb.org/newsroom/release/tech-literacy-100608.pdf
Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it Matter if I take My Writing Test on Computer? An Empirical Study of Mode Effects in NAEP. *Journal of Technology, Learning, and Assessment, 5*(2).
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.
Koretz, D. (1998). Large-scale portfolio assessments in the US: Evidence pertaining to the quality of measurement. *Assessment in Education, 5*(3), 309-334.
Kozma, R. B. (2009). Transforming Education: Assessing and Teaching 21st Century Skills. In F. Scheuermann & J. Bojornsson (Eds.), *The Transition to Computer-Based Assessment* (pp. 13-23). Ispra, Italy: European Commission. Joint Research Centre.
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, 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.
MacCann, R. (2006). The equivalence of online and traditional testing for different subpopulations and item types. *British Journal of Educational Technology, 37*(1), 79-91.

Madeja, S. S. (2004). Alternative assessment strategies for schools. *Education Policy Review, 105*(5), 3-13.
MCEETYA. (2007). *National Assessment Program – ICT Literacy Years 6 & 10 Report 2005*. Carlton South, Australia: Curriculum Corporation.
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.
Pollitt, A. (2004, June 2004). *Let's stop marking exams.* Paper presented at the International Association for Educational Assessment Conference, Philadelphia.
Ridgway, J., McCusker, S., & Pead, D. (2006). *Report 10: Literature Review of E-assessment.* Bristol, UK: Futurelab.
Ripley, M. (2009). Transformational Computer-based Testing. In F. Scheuermann & J. Bojornsson (Eds.), *The Transition to Computer-Based Assessment* (pp. 92-98). Ispra, Italy: European Commission. Joint Research Centre.
Scheuermann, F., & Bojornsson, J. (Eds.). (2009). *The Transition to Computer-Based Assessment*. Ispra, Italy: European Commission. Joint Research Centre.
Siozos, P., Palaigeorgiou, G., Triantafyllakos, G., & Despotakis, T. (2009). Computer based testing using 'digital ink': participatory design of a tablet PC based assessment application for secondary education. *Computers & Education, 52*, 811–819.