1-1-2013

Using digital representations of practical production work for summative assessment

C. Paul Newhouse
Edith Cowan University

10.1080/0969594X.2013.868341
This is an Accepted Manuscript of an article published by Taylor & Francis in Assessment in Education on 2 January 2014 as: Newhouse, C. P. (2013). Using digital representations of practical production work for summative assessment. Assessment in Education: Principles, Policy & Practice, 21(2), 205-220. Available online here

Using digital representations of practical production work for summative assessment

C. Paul Newhouse
Associate Professor, School of Education, Edith Cowan University, Perth, Western Australia

Abstract

This paper presents the findings of the first phase of a three-year study investigating the efficacy of the digitisation of creative practical work as digital portfolios for the purposes of high-stakes summative assessment. At the same time the paired comparisons method of scoring was tried as an alternative to analytical rubric-based marking because we believed that it was likely that a more holistic approach to scoring would be more appropriate. Researchers created digital representations of the practical submissions of 75 Visual Arts and 82 Design students graduating from secondary school in Western Australia. These digital portfolios were scored using the two methods with the scores compared to those officially awarded to the physical forms. It was concluded that the digital representations of the Visual Arts submissions had adequate fidelity for the purpose of awarding high-stakes scores particularly using the paired comparisons method. However the Visual Arts teachers and students were opposed to digitisation. For the Design portfolios teachers and students were supportive of digital submission but the structure of the portfolios reduced the reliability and validity of scores, particularly from analytical marking.

Keywords: digital portfolios, summative assessment, paired comparisons, Rasch modelling, Visual Arts, Design

Author Biography

Paul Newhouse, Ph.D. is the director of the Centre for Schooling and Learning Technologies (CSaLT) in the School of Education, Edith Cowan University in Perth, Western Australia. His research interests focus on using ICT to support learning in schools, particularly with regard to portable computing, assessment and technology education curriculum.

Contact Information

Correspondence concerning this article should be addressed to Paul Newhouse, School of Education, Edith Cowan University, 2 Bradford Street, Mount Lawley, 6050, Western Australia.

E-mail: p.newhouse@ecu.edu.au
Phone: +61 8 93706469
Fax: +61 8 93706397

Introduction

In Western Australia (WA) students submit a physical portfolio of artefacts and/or documents for high-stakes summative assessment at the end of some senior secondary courses. Not only are the logistics of managing thousands of often bulky materials over thousands of kilometres difficult, but so too is the reliable scoring of typically diverse forms of portfolios representing different contexts. One solution would be for students to submit digital representations of their portfolios that would allow assessors to access them on the Internet from anywhere using statistically enhanced methods of scoring. If the material to be assessed is in digital form it is more feasible to consider other methods of scoring and particularly those involving more holistic judgements. While this all seems compelling it relies on the digital representations adequately replacing the original physical forms, or as Dillon and Brown (2006) argue, the resulting digital portfolio must demonstrate adequate fidelity to gain the confidence of all stakeholders, including assessors, students and teachers.

In response we developed a three-year study that built upon the success of five-years of research focussed on using digital technologies to support performance assessment in courses with a major practical component (Newhouse, 2010), and upon collaboration with researchers in the British e-scape project (Kimbell, Wheeler, Miller, & Pollitt, 2007). One of the outcomes of this earlier research was the successful application of an online paired comparisons (sometimes referred to as comparative pairs) method of scoring digital portfolios (Newhouse & Njiru, 2009). To consider paired comparison scoring required the simultaneous consideration of digital representation of the portfolios because this method of scoring would be unmanageable on the scale required if using physical representations. This paper presents the findings from the first phase of the study that investigated the potential of replacing physical forms with digital forms of portfolios for submission for high-stakes summative assessment, and to score them using a paired comparisons method. The study used the *Visual*

Arts and *Design* senior secondary school courses in WA as examples of the different types of creative practical work that is assessed. Initially the paper sets out a rationale from the literature, then briefly explains the methodology, and finally discusses the data, analysis and findings from the first phase.

Rationale for the study

The use of physical portfolios for the assessment of practical work presents a number of key dilemmas, particularly for high-stakes purposes. Firstly, they are typically bulky, making it expensive and difficult to manage if they are to be centrally scored (Madeja, 2004; Stecher, 1998). Secondly, it is difficult to generate reliable scores due to the subjectivity of assessors and the typically varied contexts; a problem long recognised in the human judgement of creative expression (Brookhart, 2013; Koretz, 1998). For example, in the *Visual Arts* course in WA 'portfolios' (termed practical submissions) may include artistic artefacts that are 2-dimensional, 3-dimensional or digital and students may be over 1000 kilometres away from the assessment centre. Further, in the *Design* course detailed design documents are submitted to explain the development of design artefacts whereas in the *Visual Arts* course a very limited 'artist statement' is submitted. Where the assessment is summative in nature, it is critical that judgements measure performance reliably and validly. The study drew from three main fields of research: portfolio assessment, psychometrics and computer-supported assessment in terms of digital representation of creative expression and online scoring.

Portfolio assessment is not new and is regularly used for low-stakes or formative assessment purposes but its use for high-stakes summative assessment has been considerably less prevalent largely due to management and scoring difficulties (Clarke-Midura & Dede, 2010; Koretz, 1998). Portfolios are used for what Messick (1994) calls "performance-and-product assessment" (p. 14) where he distinguishes between a performance, which concerns processes and procedures, and a product that is a remaining outcome. In the *Visual Arts*

course the focus of the portfolio assessment was on the product whereas for the *Design* course the focus was on the processes and procedures. However, neither was a developmental portfolio being rather a collation of evidence at a point in time.

Psychometrics is the field of measurement of psychological attributes concerned with quantifying mental variables that are typically considered by nature to be qualitative (Barrett, 2003). It is a critical field of research for summative performance assessment, particularly in the arts where assessment necessarily relies on subjective judgements. Humphry and Heldsinger (2009) discuss this dilemma in the use of rubrics for analytical marking in performance assessment and the application of Rasch modeling. However, Pollitt (2004) calls into question the whole traditional analytical approach of summing scores on "micro-judgements" explaining that this introduces considerable error into the measurement of performance that has "harmful consequences" (p. 5). He recommends the use of holistic judgements as illustrated in the paired comparisons method, that incorporates Rasch modeling, as used in the e-scape project (Kimbell, et al., 2007).

Computer-supported assessment includes any situation in which computer technology is used to support at least part of the process of assessment whether that be students, teachers or assessors using the technology (Bull & Sharp, 2000). Typically research has focussed on the higher education sector (e.g. Brewer, 2004) and portfolio assessment has referred to physical forms, often in the arts (e.g. Madeja, 2004). As computer technology has developed into more powerful, cheaper and more flexible and integrated forms the interest in its application to problems of assessment has grown. Educators, such as McGaw (2006), have argued that with computer support and modern psychometrics, summative assessment could be better aligned with intended curriculum outcomes and preferred pedagogies. Therefore research needs to be conducted into the feasibility of using digital portfolios for assessment on complex creative tasks, particularly concerning manageability and measurement reliability

(Clarke-Midura & Dede, 2010; Ridgway, McCusker, & Pead, 2004). Dillon and Brown (2006) have addressed some of these issues and developed protocols for the use of digital portfolios in a number of areas of the creative arts. They began with the question concerning what "constitutes knowledge in the discipline" (p. 430), then consider how this "knowledge can best be represented in media" (p. 430) before determining technical requirements such as file format, size and sensory quality. Our study sought to follow these directions to address three specific questions.

1. What techniques and procedures are appropriate for the faithful digital conversion of each typical type of portfolio to support the summative assessment of student performance in the Visual Arts and Design senior secondary courses?
2. Does the paired comparison judgements method deliver reliable results when applied to digitised portfolios involving a variety of types of media and contexts?
3. Are the results of assessing the digitised portfolios consistent with assessing the original portfolios and what are the likely causes of any discrepancies?

Method

The design of the study was as an action-research evaluation involving the use of interpretive techniques with qualitative and quantitative data contributed by students, teachers, and assessors. Measures of achievement and cost were largely quantitative in nature while more qualitative data from observation, interview and survey were used to interpret the ethnographic context.

The study was conceived in two development-evaluation phases, a 'development and pilot' phase and a 'school-based implementation' phase. The first phase, the focus of this paper, was to explore the adequacy of representing the practical work in digital forms and scoring it using a paired comparisons method, and the second was to determine the feasibility

of students creating the digital representations themselves in school. An initial situation analysis by the researchers reviewed syllabus requirements and the nature and structure of the portfolios submitted. From this specifications for digitisation and the techniques and equipment required were determined. For the first phase, the sample comprised ten teachers in *Visual Arts* and six in *Design* and their 75 *Visual Arts* and 82 *Design* Year 12 students who had submitted work for external assessment.

The research team used scanners and cameras to represent the practical submission of each student in a set of digital files of various types that were then stored on servers as digital repositories. Experienced assessors were used to score the work. Interviews and questionnaires were used to elicit the perceptions and attitudes of students, teachers and assessors. Data were analysed both for each class and for the whole sample, within a feasibility framework adapted from the work of Kimbell and Wheeler (2005) and comprising the four dimensions of manageability, technical affordance, functional operation (validity and reliability of measure), and pedagogic alignment. It is appreciated that there is a tension between these dimensions, in particular as Stobart (2008) explained with a 'one-handed clock' metaphor, improvements in one dimension come at a cost to one or more of the others.

In the first phase the functional operation dimension was of paramount importance and thus the focus was on scoring and the analysis of the resulting scores. Two methods of scoring the digital representations were used, analytical and paired comparisons. In addition the scores from the official marking of the original physical submissions were obtained. Quantitative analysis of the resulting scores through correlation, scale analysis and Rasch modelling provided evidence of the relative reliability of these measures. For both methods of scoring reliability coefficients were generated using scale analysis for the analytical scores and Rasch modelling analysis for paired comparisons (Andrich, 1982). A measure of convergent validity of the assessment was investigated through a comparison with the scores

from the official scoring of the original physical submissions, augmented with the perceptions of students, teachers and assessors.

The development tasks

The research team completed a number of development tasks for the first phase starting with defining the portfolio requirements for the two courses, the digitisation specifications, assessment criteria and the design of the scoring tools. The research team reviewed the Western Australian Certificate of Education (WACE) submission requirements and agreed on a set of requirements and specifications for the digitising of the practical submissions for each course and each type of submission (e.g. Visual Arts 2-D and 3-D). The student questionnaire, teacher and assessor interview proforma were modified from a previous study.

The portfolio requirements and digitisation

Each *Design* course portfolio comprised up to 15 single-sided A3 paper pages on which students had the freedom to select examples from up to three design projects completed throughout the course. The aim was to provide evidence of knowledge and skills in a design context with an emphasis on quality not quantity. Two researchers used an A3 colour scanner to generate PDF files of these portfolios, wherever possible feeding the entire portfolio through the scanner automatically to generate a single PDF file.

The *Visual Arts* course portfolio required a resolved artwork, an artist statement, and a printed photograph of the completed artwork. There were three categories of submission each with defined constraints: Two-dimensional; Three-dimensional; and Motion and time-based. There were none of the third category in our sample. All these portfolios were stored at one large hall and so on one day teams of researchers created the digital still images and videos for our 75 submissions using SLR digital cameras and digital video cameras. For some three-dimensional work a motorised turn-table was used to assist in creating the video. Due to severe time constraints and limitations of space it was not possible to fully implement the

intended digitising procedures, however, the best attempt was made. For example, it was not possible to set up specialised lighting or backdrops, and photographs and videos could not be checked and retaken. For each submission at least one main photo (and up to 10), a photograph of the artist statement, and a short video were recorded. Later four close-ups were digitally constructed from the main photo(s) based on the decisions of an art education expert. For some 3D works a virtual reality video was also constructed. Finally, all photographs were combined in a single PDF file.

Assessment criteria and tools

Analytical marking criteria were taken from the course documentation and presented in the form of a rubric, with levels of performance described for each criterion linked to a numeric score (the criteria titles and score allocations are shown in Table 1).

<TABLE 1 HERE>

An holistic criterion was collaboratively distilled from the analytical criteria for the paired comparisons method of scoring. The holistic criteria were as follows.

Design: Judgement about performance addresses students' ability to apply elements and principles of design in recognising, analysing and solving specified design problems innovatively with consideration for a target audience and justify design decisions through experimentation and production.

Visual Arts: Judgement about performance addresses students' ability to creatively use visual language, materials and processes to skilfully communicate an innovative idea in a resolved artwork.

The scoring was facilitated by a combination of custom built and commercially available online tools that accessed the digital representations from servers via the Internet. An analytical marking tool was customised for each course, based on one developed for a previous study, using the relational database software *FileMaker Pro* (Filemaker Inc., 2007).

It was accessed through a standard web-browser and incorporated the rubrics, radio-buttons to indicate scores on each criterion, and displays of the students' digitised work. An online scoring tool called the *Adaptive Comparative Judgements System (ACJS)* developed with the *MAPS* portfolio system for the e-scape research project was used for paired comparisons scoring (Pollitt, 2012). The *ACJS*, as described by Pollitt (2012), is accessed through a standard web-browser and incorporates all of the processes of the paired comparisons method of scoring including generating the pairs for assessors to judge and allowing them to view each portfolio, recording those judgements, and applying Rasch dichotomous modelling to estimate scores and reliability coefficients. This meant that assessors only needed to judge pairs until an acceptable level of reliability was attained and they did not have to wait for others to finish.

Results

The analysis of the scores was pivotal to determining the functional operation feasibility of the portfolios. Then the qualitative data from interviews and surveys were analysed to address the other feasibility dimensions, and in particular the feasibility of digitising the particular types of portfolios. The results for each course are discussed together.

Scores from marking and judging

External assessors (two for *Design* and three for *Visual Arts)* used the analytical online tool to independently mark each student's digitised portfolio. For the paired comparisons method some of the research team, some curriculum officers, and many of the teachers in the study, were added as assessors.

The amount of time taken for analytical marking varied with an average per portfolio per assessor of 6.4 minutes for *Design* and 9.9 minutes for *Visual Arts*. For *Design* because all portfolios were of the same size the longer times were for portfolios in which it was more difficult to find the information to make judgements whereas for *Visual Arts* the longer times

were associated with work that had more components and required larger files to be downloaded. For paired comparisons judging, the *ACJS* estimated for *Design* an average of 5.6 minutes per judgement, and for *Visual Arts* 5.4 minutes per judgement. Of note for *Design* was the gradual reduction in time (5.8 mins down to 3.0 minutes per judgement) taken for each of the first eight rounds, as judges became more familiar with the tool and the material. There was then an increase in time, probably because progressively pairs provided to judges were closer in performance quality.

Analytical marking

The scores from the analytical assessors were compared with each other, and the average score between them with the official practical score (referred to as the WACE score) that used the same criteria/rubric to mark the physical portfolios. In fact the only difference between the two methods of scoring was that for the WACE there was a process of reconciling differences between the scores provided by the two assessors. Table 2 provides summaries of basic descriptive statistics and Cronbach's Alpha reliability coefficients for the scores from analytical marking. The high values for the reliability coefficients is an indication of internal consistency for the scores for each assessor and the average scores. Assessor 2 for *Design* gave a slightly wider range of scores but there was no significant difference between the mean score given by each assessor. Despite the differences between the minimum scores given by the three *Visual Arts* assessors, there was no significant difference between their average scores. For *Design* the WACE scores tended to be slightly higher (nearly 1 SD) than the scores given by our assessors while for *Visual Arts* the WACE average score was not significantly different to our assessors' average score.

<TABLE 2 HERE>

Correlation coefficients between these sets of scores are shown in Table 3 and scatter plots in Figure 1. For both courses there were only moderate correlations ($r \approx 0.5$) between the

scores allocated by our assessors using analytical marking. As a measure of the consistency between their interpretation of the marking criteria this represents a relatively low inter-rater reliability for the scoring. However, their consistency with the scores awarded for the WACE was very different for the two courses. For *Design* there were only moderate to low correlations between the scores from the our assessors and those awarded for the WACE (r=0.36 to 0.55) while for *Visual Arts* these were moderate to high (r=0.70 to 0.86). This difference is clearly seen in the scatter plots in Figure 1 and suggests that the combined judgement of the *Visual Arts* assessors were reasonably consistent with the WACE scores, while for *Design* they were not.

<TABLE 3 HERE>

<FIGURE 1 HERE>

Paired comparisons judging

The 82 scanned *Design* portfolios and the 75 digitised *Visual Arts* submissions were judged using the paired comparisons method by 9 and 14 assessors respectively using the *ACJS* online tool. For each course an initial half-day workshop was conducted to introduce the method, develop and agree on assessment criteria, and learn to use the *ACJS*. There was some discussion about the need to make a holistic judgement but keep in tension criteria related to process, technical capability and understanding of principles. Judging commenced at the workshop and then was completed over 4 weeks working from homes or workplaces.

From the beginning it was decided to stop judging when the reliability coefficient (analogous to Cronbach's Alpha (Pollitt, 2012)) was 0.95, which coincidentally occurred for both courses after the 13th round. Thus reliable sets of scores were generated. For *Design* only nine portfolios had a SE above 1.1 logits and for *Visual Arts* only eight. The mean residual was similar for all judges and close to the mean. The misfit statistic based on an Infit weighted mean square had a mean of 0.95 and 0.93 and standard deviation of 0.22 and 0.32

respectively for *Design* and *Visual Arts*, with only three and two judges respectively lying just outside one standard deviation, but still within two. Pollitt (2012) explains this misfit statistic and argues that significant misfit only occurs beyond two standard deviations from the mean.

Table 4 provides the correlation coefficients between the three sets of scores for both courses. Correlation coefficients for rankings were very similar. For *Design* the strongest correlation was between the *Pairs* scores from the digital representations and *WACE* scores from the paper-based portfolios. Possibly the combined judgements of the larger number of assessors in the *Pairs* judging was more useful than just the two analytical assessors. For *Visual Arts* there were moderate to strong correlations between all three sets of scores. The strength of the correlations with the *WACE* scores provides some evidence that the digital representations were of adequate fidelity for the purposes of external scoring.

<TABLE 4 HERE>

Differences between rankings from the two methods of scoring

There were substantial differences in ranking from the two methods of scoring for some portfolios. Initially the absolute difference between the *Analytical* rank and *Pairs* rank was calculated for each portfolio. They were compared with absolute differences between the rankings provided by each individual analytic assessors, and also between the *Pairs* rank and the ranking based on the *WACE* scores. Correlation analysis between these sets of differences was conducted to determine whether similar patterns of differences occurred between methods of scoring and representations of the portfolios. The only significant correlation was moderate ($r=0.53$, $p<0.01$) and for *Visual Arts* for difference between rankings from *Analytical* and *Pairs* scoring and between *Pairs* and *WACE*. Therefore for *Visual Arts* it was concluded that some of the difference in ranking was likely to be due to the responses of assessors to some work that may have evoked strong holistic responses. These were likely to have influenced pairwise comparison judgements but may have been moderated when

applying analytical assessment criteria. However, for both courses it was concluded that the main reason for difference in ranking was due to the relatively small sample sizes with, for example, an average difference in analytical score of 2.5 for *Visual Arts*, which is less than half a standard deviation, leading to an average difference in ranking on the analytical scores of 9.6 (nearly 13% of the 75 positions).

There were 24 portfolios in *Design* and 12 in *Visual Arts* with a difference in ranking of more than 2 standard deviations. These portfolios were reviewed in detail by asking a curriculum expert to view the work and by analysing the comments that assessors had typed into the *ACJS*. The difference in judgements between assessors for these *Design* students appeared to have been caused by differences in each assessor's priorities in judging the work. For example, Assessor 1 appeared to put more focus on the design process while Assessor 2 appeared to prefer to judge the product and visual communication skills. It was concluded that differences in ranking were partly due to characteristics of the portfolios and assessment criteria. In particular it appeared that there was too much information to be able to consistently extract what was relevant to specific assessment criteria and therefore sampling would occur leading to basing judgements on different samples of information.

Assessor perceptions

The assessors responded to a set of questions about the suitability of the digital representations of student work, the scoring processes and their perception of the quality of the portfolios.

The *Design* assessors indicated that the quality of work ranged from moderate to very good. Some criticised the presentation and layout, for example, being "too cluttered with written text" or the inclusion of content, for example, wasted space "to 'please assessors' instead of showing conceptual development, refinements to concepts". For *Visual Arts* the

assessors had a mix of opinions on the quality of the work from low, average, to above average; however all assessors agreed that there was no particularly impressive work.

The fidelity of digital representation was a critical concern for the study. The *Design* assessors generally considered that the digital portfolios represented the student work well. Most considered that moving to digital portfolios was important, especially because that would be one of the requirements in both industry and tertiary study. In contrast the *Visual Arts* assessors reported that the quality of the digital representations was poor. Some reported that the photographs were blurry and did not represent the scale, details, textures, media, and dimensions of the real work, especially the 3D works. Further, some reported that the videos were wobbly, shaky, and aside from showing an indication of the size of the artwork, did not contribute much to the perception of the work. Because the artworks were photographed and video-recorded in front of other artworks, most assessors found the background to be distracting. They were critical of the image resolution, lighting, leaning easel, and that multi-piece works did not present in a unified way. One assessor suggested that some photos reduced faults that were easier to see in 'real life'.

With regard to the experience of using the scoring tools, both sets of assessors found that the interfaces worked well and there were only a few who had problems with network speed. Two *Visual Arts* assessors suggested side-by-side viewing of student work (as was the case for *Design*) would be better for judging using the *AJCS*. For *Design* the *AJCS* scoring process was reported to be "enjoyable" and easy, with two assessors reporting they found the holistic criterion easy to use. For *Visual Arts* most of the assessors had seen some of the 'real' artwork and they considered the experience influenced their judgement. One assessor suggested that the artist statement should include more information to help "inform markers of materials and supports used, as this is very hard to discern in the 2D format".

In comparing *Analytical* and *Pairs* scoring one *Visual Arts* assessor perceived the analytical marking to be more reliable and consistent because there were "criteria to base the judgement on". Another considered comparing two artworks to be easier because there were many judges, making it more reliable. She recognised that analytical marking was still subjective, despite being based on set criteria. For *Design* the *Pairs* assessors considered that the method would increase the reliability of the scoring because of the number of assessors and judging cycles. Because there was only one holistic criterion most assessors found that it eliminated the possibility of different interpretations, discrepancies in the weighting, and the influence of personal expectations. However, the two *Design* analytical assessors had differing views as illustrated by the following quotes.

> *I would prefer analytical marking as this allows me to analyse and judge one design work at a time. This focus is more detailed and accurate - for me.*
>
> *I found the pairs marking less demanding than analytical marking. I didn't need to hold standards in my head. … My guess is that the pairs method will be the most reliable.*

Survey of students

The students completed a questionnaire consisting of closed-response and open-response items. Basic descriptive statistics were calculated for closed-response items and the three scales constructed from sets of these items (*eAssess*, *Skills* and *School Computer Use*). Data from open-response items were collated and then organised to draw out generalisations. The intention was to both get feedback on the results of digitisation in the first phase of the study and to identify relevant characteristics of the students in preparation for the second phase in which students would complete the digitisation.

Over 85% of students indicated having access to desktop computers and digital cameras at home. At school the *Design* students used computers for an average of 72 minutes per day compared with only 42 minutes for *Visual Arts* students. They indicated higher levels of skill in file management and image editing (at least 71% of *Design* and 44% of *Visual Arts*

students with excellent skills) and lower levels of skill in using web authoring and video editing software. Both groups indicated a reasonable level of computer skill with a mean on the *Skills* scale of 3.3 for *Design* students and 2.9 for *Visual Arts* students, both above the 2.5 mid-point.

Students in the *Design* course generally indicated that they had previous experience in representing their work digitally and all but 6.5% of them felt they would readily get used to the process. A mean score of 3.0 on the four-point *eAssess* scale indicated a general positive perception of the digital portfolios. However, about half disagreed with the statement, "The digital portfolio represents my design work very well", about 80% of them would have preferred an assessor to mark their original work rather than the digitised version, and 85% would have preferred to create the digital portfolio themselves. It is likely that the combination of these two responses indicates that for the vast majority they would be happy to have their digital portfolio assessed if they had digitised it themselves.

Students in the *Visual Arts* course indicated that they had little experience in representing their work digitally (44% indicated no experience) and only about 30% of the students felt they would be able to quickly adjust to the process. The mean score of 2.8 on the *eAssess* scale represented a generally positive perception of the digitised work although less so than for *Design*. However, about 72% disagreed with the statement, "The digital portfolio represents my *Visual Arts* work very well", 96% would have preferred an assessor to mark their original work, and only 46% would have preferred to create the digital portfolio themselves.

Interviews with teachers

Four of the six *Design* teachers and nine of the ten *Visual Arts* teachers provided responses to questions about the efficacy and fidelity of the digital portfolios and the responses of their students.

All of the *Design* teachers were generally positive toward the concept of digitising the portfolios for assessment. They already had their students working significantly in digital modes and they saw this as important as a way of keeping pace with industry. They believed that the portfolios should be created using computer software (e.g. by saving directly as a PDF document) rather than by scanning. They perceived several advantages including storage, ease of access, distilling a large body of work into a more refined portfolio, saving time and money, ease of management and organisation, and future use of the portfolio such as for job applications. They perceived some disadvantages including a lack of appreciation of original drawings, the lack of ability to represent more tactile designs, and that all students would need some background knowledge in graphic design. Overall they were confident in the capability of their students to produce portfolios digitally; indeed most of their students already did so.

All the *Visual Arts* teachers were opposed to the idea of using digital representations for the practical submissions. They believed that the critical attributes of artwork could not be consistently demonstrated in digitised form, in particular texture, colour, scale, impact, mounting, three dimensions, media used (e.g. photographing glass or perspex). Additionally, there were concerns about inequities in lighting, camera quality, potential use of professional photographers, and potential for manipulation of the digital images. They did see some advantages including logistics and time for transport of works for assessment, reduced chance of work being damaged, cost reduction, less restriction on work size, and that digitised work could be sent to exhibitions for selection. Four did not believe their students had the skills to adequately represent their own work digitally, and the other six felt that only some of their students would be capable of doing this. Three said that they wouldn't let students make their own digital representations because they believed the process would need to be teacher-guided.

Feasibility

The results of data analysis from the quantitative and qualitative sources were synthesised using the four dimensions of a feasibility framework adapted for the study from the e-scape project (Kimbell, et al., 2007): manageability; technical affordance; functional operation (validity and reliability); and pedagogic alignment. Results and conclusions are summarised below using the structure of this feasibility framework.

With regards to *manageability*, although scanning the paper-based portfolios for *Design* was straightforward for both courses the centralised digitisation was not feasible for system wide implementation. For the *Visual Arts* course creating the digital representations was difficult and time-consuming. However, making the digital representations available to assessors for both methods of scoring was relatively easy to accomplish. For both courses it would seem to be more feasible to digitise the work at school and submit it online. This was to be investigated in the next phase of the study. Teachers saw advantages including logistics and time for transport of works for assessment, reduced chance of work being damaged, and cost reduction. Some *Visual Arts* teachers foresaw difficulties in managing students in the creation of their own digital representations and that the students may have inadequate technical skills.

With regards to *technical affordance,* it was demonstrated that it was technically feasible to adequately represent each type of practical work in digital forms using either a scanner, or still and video cameras and the specifications developed for the study. Unfortunately for *Visual Arts* the intentions of the specifications were not realised due to logistical constraints where time and space did not permit the use of appropriate lighting, backdrops and technical photographic adjustments or virtual reality representations. Some teachers felt the quality of the representations, particularly in terms of resolution and colour

reproduction, was inadequate, however, this did not appear to noticeably influence the results of scoring.

With regards to *functional operation*, the inter-assessor agreement for analytical marking was poor, though there was evidence of good internal consistency for each assessor. With paired comparisons judging the scores showed high consistency between the judges, particularly for *Visual Arts*. In general inconsistency between assessors was probably due to a high level of subjectivity in the interpretation of the criteria for *Visual Arts* and the quantity and complexity of information for *Design*, and probably occurred equally in marking the digital and physical representations. Comparing the scoring of the digital with the physical works, there was only low to moderate correlation for *Design* but a high level of correlation for *Visual Arts*. This may have arisen partly from the assessors having seen some of the original artwork, although this was also the case for the two *Design* analytical assessors. On validity, the vast majority of teachers and students in *Design* were positive about the validity of using digital representations of practical design work provided the students created their own digital representations, but for *Visual Arts* most were negative for a range of reasons.

With regards to *pedagogic alignment* using digital representations of practical work for assessment was very consistent with intentions and practices for the *Design* course but not for *Visual Arts*. For the *Design* course most students already completed at least some of the contents of their portfolios digitally and they, and their teachers, believed that submitting paper-based portfolios was not aligned with the intentions of the course. However, about one-third of the students indicated limited experience with digital portfolios and felt they would need some time to become proficient. For the *Visual Arts* course some students and teachers perceived value in representing art digitally but not for assessment purposes. In general teachers did not believe that digital representation aligned with the purpose of practical art work and it was not part of their current teaching.

Conclusion

The first phase of the study demonstrated the affordances of relatively inexpensive and accessible digital technologies to create digital representations of students' practical work in the *Design* and *Visual Arts* courses with reasonable levels of fidelity for the purposes of summative assessment. In terms of the first research question, techniques and procedures were developed that supported the faithful digital conversion of the types of portfolios required for the two courses. However, the study identified limitations in the structure of the *Design* portfolio and the generally negative attitudes of the teachers and students in *Visual Arts* towards replacing the assessment of the physical submission with digital representations. Further, relatively standard and accessible online systems could be used to support the scoring of this work with fairly minimal maintenance requirements. This allowed the paired comparisons method of scoring to be employed that appeared to provide reliable scores for both courses and, in particular, appeared to be better suited to the *Visual Arts* work, than analytical methods of scoring where the traditional inter-rater reliability coefficients were low. Finally, the results of scoring the digitised portfolios in *Visual Arts* correlated strongly with official scores for the physical portfolios, but this was not the case for the *Design* portfolios.

Students and teachers in the *Design* course were very positive about the affordances of digital technology for summative assessment but less so if the students did not create the digital representations themselves. However, to realise these affordances the focus and structure of the portfolio may need modifying to include audiovisual representations, to reduce the amount of information and variations of layout and location, and to represent the progress of a single project. In the *Visual Arts* course students and teachers were very negative about the affordances of digital technology for summative assessment and were generally adamant that the original artwork needed to be viewed by the assessor. The external

digitisation was too cumbersome, time consuming and labour intensive and the more limited technical skills of many students may make it difficult for them to represent their artwork digitally as will be a focus of the second phase.

It is clear that the central digitisation of practical submissions in any context is probably impractical and inefficient. Therefore digitisation would only be feasible if it was conducted by the student for online submission as was planned to be the focus of the second phase of the study. To achieve this, clear technical specifications are needed to inform the digitisation process (e.g. backdrop, lighting, camera quality, file formats and size) to support technical and functional feasibility. However, consistent with assessing physical portfolios, the structure and size of a digital portfolio is critical to allow assessors to make consistent judgements, as is the structure and clarity of the assessment criteria. These recommendations have been made to the awarding body with substantial improvements made to the assessment criteria now used in both courses, and used in the second phase of our study. Finally, it was recognised that there was a growing logic for the use of digital portfolios where students tend to use digital technologies in the creative process. Despite this it is also clear that many students and teachers will need further convincing of the functional operation of these approaches to assessment and will need further technical and pedagogical support.

Acknowledgements

The study was conducted by the Centre for Schooling and Learning Technologies (CSaLT) at Edith Cowan University in collaboration with the School Curriculum and Standards Authority of Western Australia (SCASA) and supported by an Australian Research Council research grant. A team of researchers and assistants in CSaLT contributed to the first phase of the study coordinated by Dr Martin Cooper.

References

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR.20 index, and the Guttman scale response pattern. *Education Research & Perspectives, 9*(1), 95-104.

Barrett, P. T. (2003). Beyond Psychometrics: Measurement, non-quantitative structure, and applied numerics. *Journal of Managerial Psychology, 3*(18), 421-439.

Brewer, C. A. (2004). Near Real-Time Assessment of Student Learning and Understanding in Biology Courses. *Bioscience, 54*(11), 1034.

Brookhart, S. M. (2013). The use of teacher judgement for summative assessment in the USA. *Assessment in Education: Principles, Policy & Practice, 20*(1), 69-90.

Bull, J., & Sharp, D. (2000). Developments in Computer-Assisted Assessment in UK Higher Education. In R. Sims, M. O'Reilly & S. Sawkins (Eds.) conference proceedings, *Learning to Choose: Choosing to Learn* (pp. 255-260). Queensland, Australia: Australasian Society for Computers in Learning in Tertiarty Education (ASCILITE).

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

Dillon, S. C., & Brown, A. R. (2006). The art of ePortfolios: insights from the creative arts experience. In A. Jafari & C. Kaufman (Eds.), *Handbook of Research on ePortfolios* (pp. 420-433). Hershey PA: Idea Group Inc.

Filemaker Inc. (2007). Filemaker Pro 9. Santa Clara, CA: Filemaker, Inc. Retrieved from http://www.filemaker.com/products/filemaker.html

Humphry, S., & Heldsinger, S. (2009). *Do rubrics help to inform and direct teaching practice?* Paper presented at the Assessment and Student Learning: Collecting, interpreting and using data to inform teaching., Perth, Western Australia.

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.

Madeja, S. S. (2004). Alternative assessment strategies for schools. *Arts Education Policy Review, 105*(5), 3-13.

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

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

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

Newhouse, C. P., & Njiru, J. (2009). Using digital technologies and contemporary psychometrics in the assessment of performance on complex practical tasks. *Technology, Pedagogy and Education, 18*(2), 221-234.

Pollitt, A. (2004). *Let's stop marking exams.* Paper presented at the International Association for Educational Assessment Conference, Philadelphia. Retrieved June 24, 2007 from http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Conference_Papers

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

Ridgway, J., McCusker, S., & Pead, D. (2004). Literature Review of E-assessment. In K. Facer (Ed.), *Futurelab Series* (Vol. Report 10, pp. 51). Bristol, UK: NESTA Futurelab.

Stecher, B. (1998). The Local Benefits and Burdens of Large- scale Portfolio Assessment. *Assessment in Education, 5*(3), 335-351.

Stobart, G. (2008). *Testing Times, The Uses and Abuses of Assessment.* Abingdon: Routledge.

Table 1. Analytical marking criteria titles and maximum score allocations.

Design criteria	Max	Visual Arts criteria	Max
C1 Design elements and principles	6	Cr1: Creativity and innovation	6
C2 Design process	6	Cr2: Communication of ideas	5
C3 Analysis and innovation	10	Cr3: Use of visual language	12
C4 Experimentation and selectivity	10	Cr4: Use of media and/or materials	5
C5 Production knowledge and skills	10	Cr5: Use of skills and/or processes	12
C6 Communication and visual literacies	8		

Table 2. Descriptive statistics for the analytical marking of the digital and physical portfolios.

	Design				Visual Arts			
	Range	Mean	SD	α	Range	Mean	SD	α
Assessor 1	14.0 - 45.0	30.9	7.7	0.95	6.0 – 38.0	21.2	7.4	0.93
Assessor 2	12.0 - 47.0	29.7	6.8	0.95	15.0 – 38.0	24.8	5.8	0.92
Assessor 3	-	-	-	-	9.0 – 38.0	23.8	6.9	0.93
Average	14.5 - 45.0	30.3	6.3	0.96	12.3 – 37.7	23.9	6.9	0.94
WACE	15.0 - 50.0	35.2	8.2	-	10.0 – 40.0	25.3	6.3	-

Table 3. Correlations for scores from analytical marking.

	Design (N=82)				Visual Arts (N=75)				
	A1	A2	Avg.	WACE	A1	A2	A3	Avg.	WACE
Assessor 1	1	0.53	0.90	0.55	1	0.54	0.51	0.84	0.70
Assessor 2		1	0.86	0.36		1	0.56	0.82	0.75
Assessor 3			-	-			1	0.83	0.71
Average			1	0.52				1	0.86
WACE				1					1

All correlations are significantly different from 0 ($p<0.01$)

Table 4. Correlations between scores from paired comparisons, analytical and WACE scoring.

Score source		Design			Visual Arts		
		Assessors		WACE	Assessors		WACE
		Pairs	Analytical		Pairs	Analytical	
Assessors	Pairs	1	0.63	0.67	1	0.80	0.74
	Analytical		1	0.52		1	0.86
WACE				1			1

All correlations are significantly different from 0 ($p<0.01$)

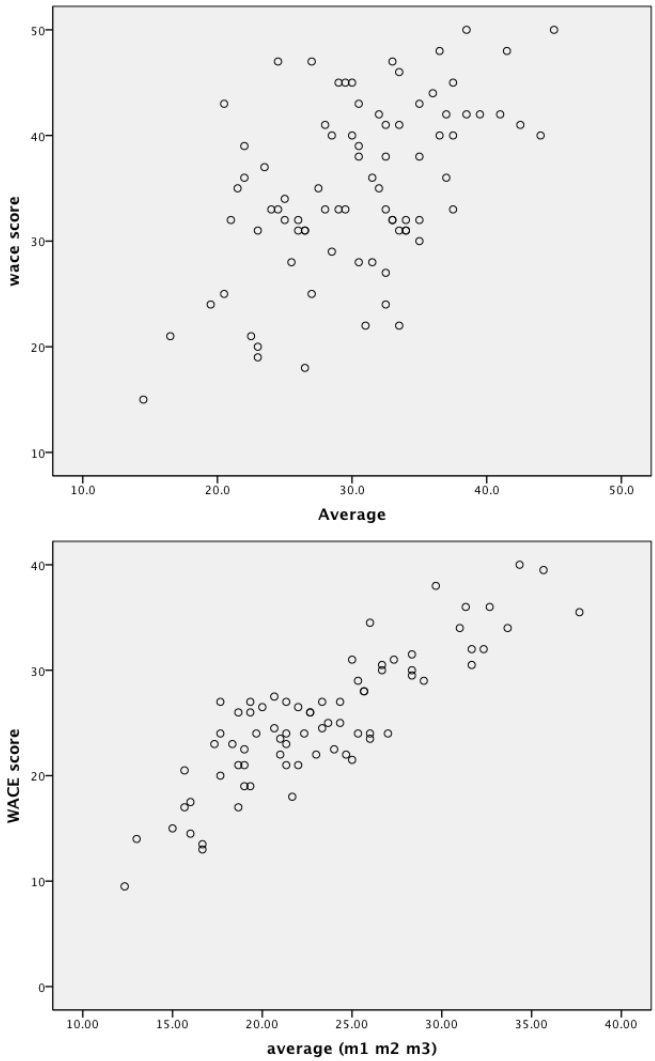

Figure 1. Scatter plots showing correlation between analytical scoring of the physical portfolios (WACE) and of the digital representations (Average) for Design (upper graph) and Visual Arts (lower graph).