Digital Forms of Performance Assessment

ACEC2008 Award Winning Paper

Abstract

In many ways assessment drives the curriculum and often pedagogy at all levels of education. In a world where it is more important what you can do with what you know rather than just what you know, it is critical that practical performance is assessed in reliable and valid ways that are also cost-effective. Increasingly the use of digital technologies to support the assessment of performance is seen as a key aspect to achieving this aim. This paper will report on the use of an online portfolio system and database linked marking tool to support a number of ways of using digital technologies to represent student work for assessment.

INTRODUCTION

This paper reports on some of the results of a small exploratory study conducted at the Centre for Schooling and Learning Technologies (CSaLT) at Edith Cowan University (ECU) in collaboration with the W.A. Curriculum Council and funded by and ECU-Industry collaborative research grant. The study commenced in August 2006, was completed by December 2007 and concerned the potential to use digital technologies to represent the output from assessment tasks in two senior secondary course, Applied Information Technology (AIT) and Engineering Studies (Eng). The study involved seven small case studies for AIT involving eight teachers and eight classes, and six small case studies for Eng involving six teachers and seven classes. The number of students involved in each case study ranged from 4 to 45 and therefore caution needs to be taken in interpreting the analysis and generalising from the results. However, the study has provided the opportunity to develop a sound methodology, including analysis techniques, and has suggested areas for further research and some guidelines for policy and practice. This paper only reports on the results related to the AIT course of study.

The focus of the study was on the use of digital technologies to ‘capture’ performance on practical tasks for the purpose of high stakes (i.e. used to determine future opportunities) summative assessment [Note: This may include representations of student ‘outputs’ or performance processes]. The purpose was to explore this potential so that such performances could be included in the assessment of senior secondary courses to increase the authenticity of the assessment in these courses. It was recognised that there was little assessment of such performances currently due to factors related to cost, validity and reliability. However, what is assessed is critical because students tend to focus on, and be motivated by these sections of the curriculum, and teachers tend to ‘teach to the test’. Further, educators are accountable to society for the outcomes of the use of resources in education, and our society was increasingly expecting that students should demonstrate practical performance not just theoretical understanding. Finally, students are more likely to experience deep learning through complex performance. Therefore as McGaw explains, this places a responsibility on education authorities to consider strategies to increase the assessment of performance on practical tasks.

“If tests designed to measure key learning in schools ignore some key areas because they are harder to measure and attention to those areas by teachers and schools is then reduced, then those responsible for the tests bear some responsibility for that” (McGaw, 2006 p.3).

In Western Australia for almost all courses students are externally assessed using traditional methods employing predominantly paper and pen technologies, however, clearly in most cases performance on practical tasks cannot be assessed on paper. Many educational researchers argue that traditional assessment only measures knowledge of basic facts and procedures but fails to assess learning processes and higher-order thinking (decision-making, reflection, reasoning and problem solving). The only advantages that paper-based exams have are low cost and ease of authentication. In the developed world a very small proportion of tasks are done on paper.

Performance-based assessment is not new. Oral and laboratory examinations have been used in European schools and Universities for over a century. In many industries performance-based assessment approaches are used (e.g. pilots). In many high-stakes courses in developed countries performance is, and has been, assessed using observation, interview, portfolio or recording (e.g. USA, UK, Denmark). For example, a recent review of assessment methods in medical education (Norcini & McKinley, 2007) outlines performance-based assessment of clinical, communications and professional skills using observations, recordings and computer-based simulations. In W.A. there has been a history of performance-based assessment in some courses in the Arts. However, the use of performance-based assessment in high-stakes courses has been limited by the costs involved in collecting the evidence of performance and difficulties in ensuring reliable and valid results. Recent advances in psychometric methods and improvements in digital technologies provide tools to assess a variety of performance relatively cost-effectively (McGaw, 2006).
Any approach or strategy will not be perfect and will require compromises and consideration of the following questions. [Note: Although this paper does not specifically address these questions they informed the decisions and data analysis for the major study.]

1. What skills or knowledge are best demonstrated through practical performance?
2. What are the critical components of that practical performance?
3. Why can’t those components be demonstrated on paper?
4. What alternative representations other than paper could be used?
5. What level of compromise in reliability, authentication and cost is acceptable in preference to NOT assessing the performance at all?

A ground-breaking study aimed at assessing performance, titled e-scape, is being conducted by the Technology Education Research Unit (TERU) at Goldsmiths College, University of London (Kimbell, Wheeler, Miller, & Pollitt, 2007), built upon many years of work on improving assessment in the design and technology curriculum (Kimbell, 2004). e-scape combines three innovations in the assessment of practical performance by representing student work entirely in digital form, collating this work using an online repository and marking it using a comparative pairs judgements technique. Their results have been encouraging with student work being assessed on an interval scale with a reliability coefficient (Rasch Separation Index that is equivalent to a Cronbach Alpha) of 0.93. The study has continued into 2007 and 2008 with an emphasis on involving a number of subject disciplines and scaling the approach for potential nation-wide implementation.

Method

The AIT component of the study involved eight teachers and their upper secondary students in seven schools. Each school was treated as a separate case with a different digital form of assessment used in each school. A list of the cases, sample sizes and brief description of the form of assessment is provided in Table 1.

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

Output from student work on the assessment tasks was collected in digital form and placed in the online digital repository to be available to the assessors. The assessment task could not be standardized and thus was quite different for each school/class. Thus there were differences in parameters such as time constraints, level of supervision, level of scripting for implementation, minimum requirements for equipment/resource availability, and types of data to collect. The students’ work was assessed by at least two external assessors and for one case a panel of five assessors, in addition to the classroom teacher, using a set of criteria developed for each assessment task. For each case a mark and ranking were created for each assessor and a consensus mark and ranking. These were compared using a test of correlation. For the case involving a panel of five assessors a comparative pairs method of marking was used to generate an overall mark and ranking for each student. This was done in addition to the standard two external assessor strategy.

Teachers provided achievement data for their students for the assessment task and for other school-based assessments. These data were provided as scores, percentages, ranks or grades. For each case these data were used to create a ‘teacher mark’ and ‘teacher ranking’ for the students. Where grades were provided these were converted to marks using conversion tables developed by the researchers for that case. The teacher mark and teacher ranking were compared with the external assessors marks and rankings using a correlation test.

Interviews were used to elicit the experiences of students in completing assessment tasks, teachers in supporting them in these, and assessors in marking the student work to address the manageability, functional and pedagogic dimensions of feasibility. A stratified sample of students in each class was interviewed as a group, and each teacher, as soon after completion of the assessment tasks as possible. The student interviews were transcribed and then summarised in the case study reports so that themes could be identified. The data from the teacher and assessor interviews were summarised and included in the case study reports.

Observations of each class of students in the process of completing the assessment task occurred using a structured approach to address the manageability dimension of feasibility. These data assisted in interpreting results from other data particularly in terms of the constraints associated with the realities of conducting these assessments in schools. Notes were written or recorded during observation periods and were verified by the participating teacher as soon as possible after the observation. A summary was included in the case study reports.

Questionnaires were employed to collect data from students on their characteristics, including perceived level of ICT skills and experience, and their experience of the assessment task. These data

<table>
<thead>
<tr>
<th>Case</th>
<th>Course</th>
<th># Teachers</th>
<th># Students</th>
<th>Digital Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>AIT: Year 11 (A/B)</td>
<td>1</td>
<td>15</td>
<td>Portfolio - slideshow</td>
</tr>
<tr>
<td>CA</td>
<td>AIT: Year 11 (A/B)</td>
<td>1</td>
<td>10</td>
<td>Exam (Production) – web page</td>
</tr>
<tr>
<td>IA</td>
<td>AIT: Year 12 Interactive Media</td>
<td>1</td>
<td>15</td>
<td>Portfolio – Flash movie</td>
</tr>
<tr>
<td>MA</td>
<td>AIT: Year 12 Industry IT</td>
<td>1</td>
<td>9</td>
<td>Exam (Production)– video</td>
</tr>
<tr>
<td>PA</td>
<td>AIT: Year 11 (A/B)</td>
<td>2</td>
<td>17</td>
<td>Portfolio – various formats</td>
</tr>
<tr>
<td>WA</td>
<td>AIT: Year 11 (A/B)</td>
<td>1</td>
<td>10</td>
<td>Portfolio – web page</td>
</tr>
<tr>
<td>XA</td>
<td>AIT: Year 11 (A/B)</td>
<td>1</td>
<td>48</td>
<td>Exam (Tasks) - poster</td>
</tr>
</tbody>
</table>
were used to address the manageability dimension of feasibility. The questionnaires were administered to all students from the sample classes as soon after the assessment tasks as possible. The data were entered into a spreadsheet and SPSS. Descriptive and frequency statistics were generated to include in the case study reports.

These quantitative and qualitative data for each class were compiled into case studies with a report generated for each case study. Each case study report was sent to the teacher(s) involved for validation and some interpretation. Triangulation of data types and sources enhanced the credibility of findings. Validity and authenticity of the assessment tasks was judged in relation to three criteria: how well the performance of students matches the curriculum outcomes; the extent to which the method of representing performance is authentic; and, whether the task and context are meaningful and relevant to students and community practice. The outcomes from each case study were then summarised according to the dimensions of feasibility so that the data could be analysed using a constant comparative approach looking for themes, trends and developing rich descriptive accounts (Patton, 1990). Data were coded according to emergent themes. Themes were constantly compared with emergent categories to establish a best fit with the data set.

Fundamentally this project investigated the use of digital forms of representation of student practical performance for summative assessment, whether the student created digital files or their performance was recorded in digital format by filming, photographing, audio recording or scanning. The overall study investigated three digital forms of assessment: Extended Production Exam, Focussed Performance Tasks, and Reflective Portfolio.

**Extended Production Exam**: the completion, under exam conditions, of one common assessment task that incorporates a full set of processes (e.g. design process, scientific investigation) and centres on one major scenario for which students are required to submit all output in digital formats.

**Focussed Performance Tasks**: a range of focussed practical tasks that students submit in digital formats after working under exam conditions. This may be at a computer or may be a digital recording of their performances.

**Reflective Process Portfolio**: a collection of digital artefacts of work output completed by the student, according to specified parameters such as form, structure, and range of samples required. These will require some reflective commentary by students.

**Data Analysis**

A range of types of quantitative and qualitative data was collected at each school contributing to each case study. These data were analysed, to address the research questions, within a feasibility framework of the four dimensions described in Table 2.

### Table 2 Descriptions of the dimensions of feasibility.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manageability</td>
<td>Concerning making a digital form of assessment do-able in typical classrooms with the normal range of students.</td>
</tr>
<tr>
<td>Technical</td>
<td>Concerning the extent to which existing technologies can be adapted for assessment purposes within course requirements.</td>
</tr>
</tbody>
</table>

---

In addition conclusions were drawn with respect to the constraints and benefits of the form of assessment used in each case study. A summary of findings from the AIT case studies was compiled based on the Feasibility Framework and including a summary of the constraints and benefits of the form of assessment used in the case study. The **Functional** dimension was divided into findings regarding validity and those regarding reliability. Validity was analysed by considering: (1) how well the performance of students matches the curriculum outcomes; (2) the extent to which the method of representing performance was authentic; and, (3) whether the task and context were meaningful and relevant to students and community practice. A summary of findings from the combination of cases is presented in Table 3 (over page).

**RESULTS**

The results will be discussed in terms of conclusions about exams and portfolios, and conclusions about different methods of marking.

**Conclusions About Exams**

Two of the exams were production exams and one was a performance task exam. The performance task exam was short, easy to manage and mark but was limited in scope for assessment of student capability. The longer production exams provided scope for assessment but lacked adequate scaffolding to ensure students provided adequate explanation of their intentions in design and production. The degree of structure was an issue where too little structure risked students not including enough to assess and too much limiting the opportunity to demonstrate higher level achievement. In all three cases greater control over student time use would have increased the quality of evidence of achievement. That is, students needed a clear allocation of time to annotation and documentation, and to production. Another issue was the level of access students were permitted to information and resources. In two cases students had unlimited access that in one case distracted them from completing the requirements while in the other case was only used to access graphic images. In the third cases students only had access to software help files and their exam folders that probably reduced the effectiveness of their designs.

The three cases investigated were:

1. Production of a poster (50 minutes)
2. Production of a web-page with responses to some questions (3 hours)
3. Production of a video with technology process documentation (2 weeks)

The following initial conclusions were made from this analysis:

- All three were implemented successfully with the poster production the least problematic. In only one case did a computer crash (video production) with the student losing some work.
There is a fine balance between providing adequate structured scaffolding and providing adequate flexibility to allow demonstration of achievement. The Poster exam had very little scaffolding and thus marking required some 'guesswork' on intention. The video exam had a lot of scaffolding that constrained students to the extent that none of them entirely finished.

Marking by ‘adding up marks for items’ results in significantly different rankings than using standards-based rubrics. The latter provide reliable results well correlated with comparative pairs marking rankings.

**Conclusions About Portfolios**

The only set of portfolios that was relatively easy to mark was template based and in a web format. Where students were given considerable choice in either format/software or structure/style marking became more difficult and unreliable. The teacher’s knowledge of the student and other work completed by the student was likely to influence results compared with those from the assessors. It was critical that students did not have to spend a lot of time learning to use the construction software, however, this limited the pedagogic value. The degree of structure was also an issue where too little structure risked students not including enough to assess and too much limiting the opportunity to demonstrate higher level achievement. However, it appeared to be better to err on the side of too much structure.

The four cases investigated were:

1. Template-based web-site including 4 outcomes and 4 tasks
2. Tasks-based and coordinated through Powerpoint (SIDE module)
3. Skills-based in Flash with links to other tasks
4. Outcomes-based in student chosen format (e.g. Word, Powerpoint, Flash, website)

The following initial conclusions were made from this analysis.

- Template-based website is easiest to collect and to mark. All files in one folder, including an index page, then zipped and uploaded into MAPS. As well as providing an interface and means of navigation it is likely that a multi-page structure is preferable to a tagged single page.

### Table 3 Conclusions based on the dimensions of feasibility.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Manageability</strong></td>
<td>There were three case studies involving 'practical exams' with two (1 and 3 hour exams) providing no technical or logistical difficulties and in one there were students who lost work and there were logistical difficulties in managing student time over a two-week period. Of the four case studies involving portfolios in only one case were the portfolios relatively easy to access and mark. This case involved all students using a standard web-based template. The other cases were difficult either due to the range of software used, the quantity of items required, a lack of time for completion, a lack of supporting documentation, or in one case a lack of student experience and ICT capability.</td>
</tr>
<tr>
<td><strong>Technical</strong></td>
<td>Student work was collected either on a class DVD or USB drives and submitted either to a server or directly to the USB drives. Some degree of maintenance was required for all four cases involving portfolios with the least for the template-based web format. The three cases of practical exams typically only required conversion to PDF. All work in PDF, web or Flash format was easily accessed by markers.</td>
</tr>
<tr>
<td><strong>Functional</strong></td>
<td>The two external assessors used a marking rubric developed from a standards-based framework. This resulted in highly correlated marks and rankings in all cases. The comparative pairs method of marking was successfully implemented for a short performance exam with resulting highly reliable scores but questions remain concerning the validity of results with variations for particular students when cut scores were applied compared with the standards-based assessment. In only two cases (2 week production exam and low quality portfolios) were the ranking of students highly correlated between the teacher and assessors, in one case despite differences in ‘marks’. In all cases there were significant differences in the marks produced by the teacher and the assessors. In three cases this was due to the use of different outcomes and in two cases due to the use of analytic numerical marking method. In two cases differences were likely to be due to the teacher having access to more. In all cases inadequate documentation and annotation by students limited achievement to L5 or below. Further this led to differences between external assessors due to interpretation of student intention. In most cases increased scaffolding and control of time use may have partly addressed this issue. In two of the 'exams' and two of the portfolios the structure of the tasks limited the opportunity to demonstrate above L4. In five of the cases a lack of time limited the opportunity to demonstrate above L4.</td>
</tr>
<tr>
<td><strong>Pedagogic</strong></td>
<td>Typically students liked doing practical work and disliked documenting work or answering question. In at least three cases students liked learning ICT skills through the assessment. Most students demonstrated L4 work, in some cases many did not, and in some cases a few demonstrated L5 work. In general the quality of work was low probably largely as a result of the capability of the students.</td>
</tr>
</tbody>
</table>
Portfolios in Flash are easy to manage when published for the web. Once again all linked files need to be zipped within one folder.

Tasks-based portfolio was difficult to manage and mark, particularly as no student had successfully linked all the required items to a slideshow.

Insisting on a particular structure given in a template with requirement for annotations is easiest to mark. Few students provided adequate annotations for any work and would need more prompting.

Limiting file formats is essential and size of media files. For example, where possible convert files to HTML, PDF, JPG, GIF or AVI. If a slideshow was provided in PDF format then in the annotation the student could draw attention to a menu structure (may not be necessary to actually try the slideshow). Alternatively an animated slide could be recorded as a WAV file.

Marking using standards-based rubrics provides reliable rankings.

Conclusions About Methods of Marking

The standards-based method of marking using a rubric was successfully implemented with the support of a digital marking tool and an online portfolio system (MAPS) with the marks from the two assessors, and resulting rankings, highly correlated in all cases but rarely highly correlated to the teacher marks. For all cases there were only a few students whose work was judged differently by the two assessors and required a consensus meeting to resolve. Correlation coefficients for all the cases are reported in Table 4.

The comparative pairs method of marking was successfully implemented for a simple short production exam (Case XA) with resulting highly reliable scores but questions remain concerning the validity of results with variations for particular students compared with the standards-based assessment. Assessors took between 2 and 3 hours to make the comparisons for 48 scripts although it was determined that the number of comparisons may have been able to be reduced by half. It is likely that the validity may have been improved through increased training for the assessors and through providing a note-taking facility with the digital marking tool.

CONCLUSION

Overall it was concluded that all the digital forms of assessment involved in the study were implemented with no significant technical difficulties with almost all students responding positively. For the three forms of exams almost all students were able to complete the requirements in the time permitted. For the portfolios in most cases students were not able to complete the requirements in the time permitted but this was largely because the data had to be collected too early in the year. In all cases capturing the work of the students was relatively straightforward and then converting all output to files for uploading into the MAPS portfolio system. While this was an interesting result it clearly is an area that needs considerably more research. However, overall the results provide a basis on which to further investigate the use of digital forms of assessment for high stakes purposes.

REFERENCES


ACKNOWLEDGEMENT

The theory discussed in this paper and the research upon which it is based are as a result of the work of a research team organised by the Centre for Schooling and Learning Technologies at Edith Cowan University. The team was led by Paul Newhouse and John Williams and includes researchers Jeremy Pagram, and Alistair Campbell, project manager Lorraine Kershaw, and a number of research assistants.