Using Pairwise Comparisons in the Online Social Moderation of Performance Assessment

C. Paul Newhouse
School of Education, Edith Cowan University
p.newhouse@ecu.edu.au

Pina Tarricone
School of Education, Edith Cowan University
p.tarricone@ecu.edu.au

ABSTRACT

Assessing the performance of a student involves some form of judgement, and where more than one assessor is involved this usually requires some form of moderation to ensure consistent and fair results. Often this involves meetings or communication between assessors, which is referred to as social moderation. This paper reports on a study that investigated the use of online technologies to support a form of social moderation of artworks submitted for assessment in a senior secondary school course in Western Australia. Online systems were used to facilitate communications and provide access to digital representations of the submissions along with assessment tools. In particular a pairwise comparison judging online tool was used. This approach to social moderation was tested in a realistic context involving a sample of 12 teachers from rural schools for whom face-to-face meetings would be difficult. The aim was to investigate whether the use of these online systems would support good moderation outcomes and valuable professional learning for those involved. The study found that this approach to online social moderation was feasible, and participants perceived that it had improved the consistency of their judgements because they had developed an improved understanding of the assessment criteria and standard of work. However, analysis of scores and reliability data suggested some were not adequately consistent, and it was likely that this was due to their inexperience in assessing such work. Therefore some changes to the processes of this form of online social moderation were recommended.

Keywords

Online moderation; computer-supported assessment; online communications

1. INTRODUCTION

In every formal learning setting some form of assessment is used to determine the achievement of students. Typically one of the key sets of processes concern judging the student's performance on the assessment, often to generate a score or grade. Where the assessment is high-stakes there are often many assessors involved and there is the need for some form of moderation to ensure the outcome is valid, reliable and fair. Traditionally this has used either statistical methods or face-to-face meetings between assessors to reconcile judgements. Such meetings often present logistical difficulties and challenges to generate reliable scores, particularly using analytical scoring using methods such as rubrics. It is likely that online communication and database systems could be used to address these problems. Therefore, we set out to investigate this potential in the final phase of a three-year project into the use of digitized portfolios of creative work for high-stakes summative assessment in senior secondary schooling.

The aim was to focus on assessments that involved some form of practical performance where judgement would necessarily be highly subjective. Therefore, the course used was Visual Arts in the final year of secondary schooling. Online systems were used to enable assessors from any location to be involved in scoring and moderation processes. The use of online systems required that the performance of students on the assessment were represented digitally. From previous research [12; 13] we had found that the method of pairwise comparison (sometimes referred to as comparative pairs or comparative judgement) provided reliable scores where judgement of performance was highly subjective and online systems could be used to facilitate this method. We set out to use this method supported by online communication systems to facilitate an online social moderation exercise in Visual Arts. This paper now introduces online social moderation, then summarises the method for the study, followed by a discussion of the main findings concerning moderation.

2. ONLINE SOCIAL MODERATION AND PAIRWISE COMPARISON

Where assessment outcomes rely on the judgements of assessors some form of moderation is usually applied, particularly for higher-stakes instances [5]. Often this will involve communication between assessors to arrive at a consensus outcome (e.g. score or grade). This social moderation approach has been used for many years to improve the reliability and validity of the outcomes of assessment. This has particularly been the case for more subjective judgements where assessors have difficulty with consistency. For example, a study by Van der Schaaf et al. [17] investigated the reliability and validity of judgements made by teachers using a set of assessment criteria in portfolio assessment. Their concern was that teachers would vary in their interpretation of the criteria. They found that teachers were more likely to apply criteria consistently, and reflect on their judgements, where they communicated their judgements with others rather than if they were not involved in a social moderation process.

Traditionally social moderation has involved face-to-face meetings to review student work and assessments, but this is logistically difficult [11]. Replacing these with online meetings and access to student work and assessor tools online should improve the feasibility of social moderation. Such an approach is termed online social moderation and may also provide professional learning for participants through a 'community of

WWW 2017 Companion, April 3-7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054154

practice' that is able to share expertise and develop better understandings of standards and assessment [2; 7; 16; 18]. Such online social moderation may use synchronous (e.g. conferencing) and/or asynchronous (e.g. email, scoring tools) online systems allowing teachers to be engaged no matter where they are located [2; 18].

The main aim of social moderation is to enhance the consistency of judgement of the standard of student work on an assessment. The purpose of facilitating social moderation with online systems for communication and data handling is to improve efficiency and participation. The benefits of online social moderation have been discussed but there has been little implementation or research [1; 2]. However, Adie, Klenowski and Wyatt-Smith [2] conducted research in Queensland with 50 teachers from 21 disparate rural schools. In this research moderation meetings were conducted using the WebEx video-conferencing system and the telephone. The findings were that a wider participation was encouraged; the consistency of judgements of standards improved and teacher understanding was enhanced. This added to the findings from an earlier study by Klenowski and Wyatt-Smith [10] and also found that participation in these moderation exercises helped some teachers adjust their teaching to better inform students of assessment.

There are many ways in which online social moderation could be constructed. For example, teachers could score their students' work using a rubric and then use online communication systems to share the work and their judgements with other teachers. However, for judgements that are necessarily highly subjective it is difficult to justify scores with reference to an absolute description such as in a rubric. It is readily argued that it is far easier to make consistent judgements and justify them where a comparison between two pieces of work is being made. In fact comparison is fundamental to all measurement, including educational assessment [3; 15]. The latent nature of ability means that comparisons cannot be deterministic but are probabilistic. Measurement of ability relies upon comparison to infer thresholds and form an interval scale. This is the rationale for considering the use of the pairwise comparison method of judging to score work that is highly subjective [6]. This method involves multiple assessors being allocated multiple pairs of student work to adjudge the better of each pair based on an agreed holistic criterion [14]. The results of these decisions are analysed using a dichotomous Rasch model to generate scores for each piece of work on an interval scale, along with measures of the reliability of those scores.

The pairwise comparisons method has only been practical to use for large samples (>30) with the availability of online systems such as the Adaptive Comparative Judgement System (ACJS) [14] and the Pair-Wise Web Software [8]. Therefore, while for over a decade the method has been used on small samples for standards checking it is only relatively recently that this has been extended to trials on larger samples for ranking or scoring student performance on assessments. Some of the best-known examples are associated with the eScape project in the United Kingdom [9]. This project has demonstrated advantages of using the pairwise comparisons method for various types of performance, particularly where holistic judgements can be made based on digital representation of performance. They have found that the resulting scores are associated with high levels of reliability and assessors can readily be trained to implement the method.

It therefore is reasonable to suggest that it offers potential to include in online social moderation. However, there has been little use of pairwise comparison or these systems to support social moderation. Our study set out to investigate this approach to moderation with the aims of improving the reliability of scores and the knowledge and understanding of teachers for the assessment criteria and standards.

3. METHOD FOR THE STUDY

This paper reports on aspects of the final phase of a three-year study that was conducted at the Centre for Schooling and Learning Technologies (CSaLT) at Edith Cowan University in collaboration with the School Curriculum and Standards Authority (SCSA) of Western Australia and supported by an Australian Research Council (ARC) Linkage research grant. The study sought to investigate the use of digitised portfolios of creative work for summative assessment. In the final third phase of the study the focus was on using online systems to support an approach to social moderation of judgements of the performance represented by the digitised portfolios. The overall research design of the study and results from the first two phases, in 2012 and 2013 respectively, have been previously reported [12; 13]. This paper provides some background information, such as the context and digitization of the portfolios, which is needed to make sense of the findings from the final phase of the study in 2014, specifically related to online social moderation.

3.1 Context for the study

The study was set in Western Australia (WA) in the senior secondary courses of Visual Arts and Design that were part of the WA Certificate of Education (WACE) and were externally assessed for tertiary entrance. For the final phase of the study only the Visual Arts course was involved and therefore only this context is now described. For this course a practical performance assessment was used to contribute to a tertiary entrance score and therefore it was a very high-stakes assessment. Students who were studying in Year 12 of the Visual Arts course were required to be submitted for external assessment a resolved (finished) artwork (e.g. painting, sculpture, drawing, and photographs), an artist statement, and a printed photograph of the completed artwork. The artwork could be classified as Two-dimensional; Three-dimensional; and Motion and time-based. Each of these categories had specified constraints such as size. There were none of the third category in our samples, most were 2D and some were 3D. These artworks were submitted to a central location in Perth WA to be assessed by a team of expert teachers with typically each student's work judged by two assessors. The two main problems associated with these processes were that the resulting scores were likely to be unreliable, and the logistics of transporting artworks often thousands of kilometres and gathering the assessors at the central location was exacting. Therefore, we proposed an alternative approach where only digitised representations of the artworks would be uploaded to a central server, and assessors would make their judgements online from home or school. To address the judging reliability problem we proposed to use the pairwise comparison, also known as the comparative judgement method.

In the first two phases we tested whether the artworks could be adequately represented in digital forms, whether assessors could use online tools to access and judge these artworks, and whether students could create and upload these digital forms themselves. In the final phase we investigated the use of online systems to facilitate social moderation to generate reliable scores and

provide professional learning for assessors. In particular due to the expanse of WA we wanted to demonstrate that this could be achieved no matter where the assessors resided.

3.2 Portfolio digitisation

All portfolios, including those in our Phase 1 sample, were sent by the students, or their teachers, to a central location that was a large hall in a suburb of Perth. Our research team was permitted one day to access the portfolios to create the digitized representations of the 75 submissions using SLR digital cameras and digital video cameras. For some 3D works the video was recorded using a motorised turntable. Initially we had to locate our sample of artworks from amongst the thousands of submissions.

Due to the constraints of time and space it was not possible to fully implement the intended digitising procedures (refer to Table 1), however, the best attempt was made for each portfolio. We were not able to use additional lighting or backdrops and often there was no time to check and retake photographs or videos. However, for each of the 75 submissions between 1 and 10 main photographs were taken, along with a photograph of the artist statement, and a short video. At a later date we digitally constructed four close-up images from the main photo(s), under the guidance of an art education expert. Then these images along with the originals were combined in a single PDF file. In addition for some 3D works an animated virtual video was constructed. For Phase 2 another sample of students in Year 11 used similar specifications as those in Table 1 (optical close-ups instead of digitally constructed) to digitise their own artwork and upload it to an online repository.

Table 1. Intended digitising specifications

Type	Requirement	File
2 D	Photo of 'Artist Statement'	JPG
	Full size photo (Hi res 300dpi) with a matchbox included for size comparison. 72 dpi adequate for on screen viewing.	JPG
	4 x close ups – digitally extracted from main photo	JPG
	All photos combined	PDF
	HD Video (pan & zoom) - 10 secs	AVI
	Photo of proposed installation photo if provided.	JPG
3 D	Photo of 'Artist Statement'	JPG
	Full size photo + size object such as a match-box	JPG
	4 x close ups - extracted from main photo	JPG
	At least 4 x angle photos (left, right, top, bottom)	JPG
	All photos combined	PDF
	HD Video (pan & zoom) - 10 secs	AVI
	3-D Animation for selected works	AVI

3.3 Assessment criteria and tools

The criteria used for analytical marking were those used for officially scoring the art submissions, as laid out in the course documentation. They were presented in the form of a rubric, with each criterion allocated a maximum score with score-points described in terms of required performance. The criteria titles were (maximum score): Creativity and innovation (6); Communication of ideas (5); Use of visual language (12); Use of media and/or materials (5); and Use of skills and/or processes (12).

For pairwise comparison judging, a single holistic criterion was distilled from the analytical criteria in a consensus meeting with the assessors in the first phase of the study. The holistic criterion for Visual Arts was,

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 by analytical marking and pairwise judgements was done using online tools accessing the digital portfolio files from servers. A custom built analytical marking tool using Filemaker Pro [4] was adapted from a previous study to allow assessors to view the portfolio files and use a rubric to score them in a standard Internet browser (see Figure 1). The assessor clicked on the 'Exam Files' buttons in the top right to view the digital representations of the student's work and used the radio buttons in the rubric on the left to record their scores.

The pairwise comparisons method was facilitated with an online scoring tool, the Adaptive Comparative Judgements System (ACJS), developed with the MAPS portfolio system for the e-scape research project [14]. It is termed adaptive because the pairs of portfolios to be judged are generated dynamically based on the results of previous judgements, rather than all the pairs being generated at the beginning. Example user interface screens are provided in Figure 2. Assessors accessed the tool through a standard Internet browser, logged in and then viewed a pair of portfolios by clicking on the 'A' and 'B' buttons on the toolbar, then to record their judgement they clicked on 'Compare', the interface box at the bottom right popped up to allow them to indicate their selection of the 'winner' and type in any explanatory comment. The data collected from this system is automatically fed through a Rasch measurement dichotomous model to estimate scores and reliability coefficients. These are provided as online reports by ACJS, some of which can be downloaded as spreadsheets.

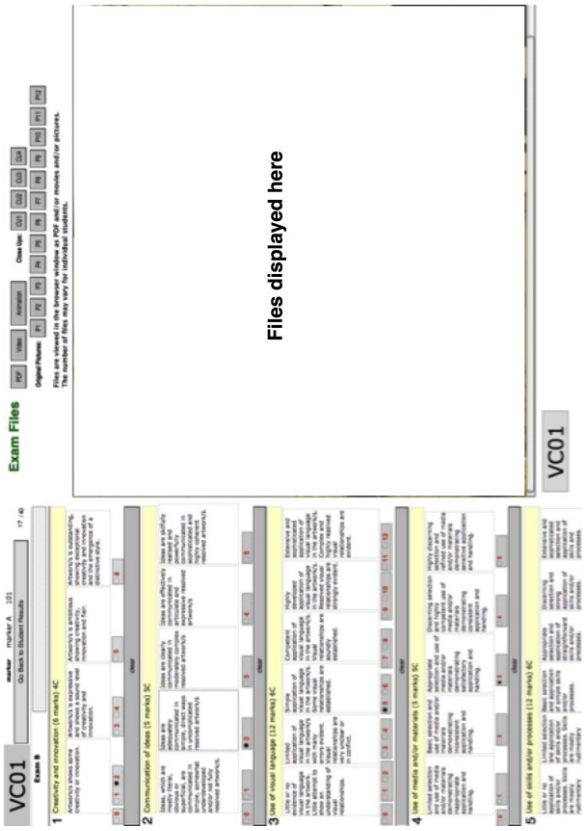

Figure 1: Interface for analytical marking tool

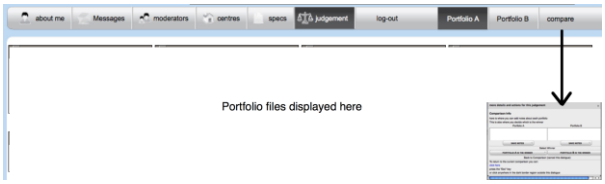

Figure 2: Interface for ACJS tool

3.4 The first two phases of the study

The first two phases of the study demonstrated firstly that the Visual Arts submissions could be digitised with adequate fidelity for the purposes of assessment, and secondly that in general students were able to digitise their own work and submit it online. This could all be achieved using relatively inexpensive and accessible technologies. Further, we demonstrated that online systems could be used to support the judging or scoring of these portfolios with minimal maintenance. This allowed investigation of the pairwise comparisons method of judging that was found to provide reliable scores. In fact given the generally low inter-rater reliability coefficients from analytical marking the pairwise method appeared to be well suited to the traditionally highly subjective nature of artworks. In addition, there was a strong correlation between these scores and the official scores for the physical submissions (this was not the case for the other course we investigated; Design).

However, we did identify that the teachers and students generally held negative attitudes and perceptions towards replacing the physical submission with digital representations. In addition, the external digitisation in Phase 1 was impractical and inefficient and it was determined that the only viable option was for students to digitise their own work. This was investigated in the second phase of the study where students followed a detailed set of technical specifications (e.g. backdrop, lighting, camera quality, file formats and size). Although this was found to be feasible it was still the case that most students and teachers were not convinced of the validity of these approaches to assessment.

3.5 Sample and procedures for the third phase

The sample for the third phase was 12 Visual Art teachers from rural schools in WA. However, the work to be assessed was the 75 digitised submissions from the first phase of the study. These files were stored on a server for analytical marking using a custom-built online tool, and uploaded into the ACJS for pairwise comparison judgement [12]. The aim of the third phase was to support these teachers, either from their schools or homes, to use these online technologies to participate in an approach to social moderation over a period of weeks. The plan for social moderation of the Visual Arts digital portfolios followed a sequence of four stages.

Stage 1. Analytical marking of a stratified sample of portfolios. Each teacher independently used an online custom-built Filemaker Pro database system tool to score the same sample of 10 portfolios using an analytical marking rubric. The intention of this exercise was to familiarise them with the assessment criteria and the range of quality of the portfolios. Therefore, the sample of portfolios had been selected to represent this range, as determined by the ranking from scores in the first phase of the study. Assessors were supported by a set of instructions and access via email and phone to members of the research team.

Stage 2. Online meeting: making pairwise judgements and using the ACJS tool. A synchronous online meeting was set up using the Adobe Connect video-conferencing system. All teachers joined from their homes or schools. The intention of this meeting was to review the 10 portfolios used in the first stage, introduce the concept of pairwise comparative judgement, and introduce the use and operation of the ACJS. At the end of the meeting we shared our screen with participants so that the group could discuss some judgements for the first few pairs of portfolios. Using chat and audio conferencing each participant could explain the basis on which they would make a judgement of the winning portfolio and a vote was taken to show the balance of judgements.

Stage 3. Pairwise comparative judgements of all portfolios. After the online meetings each assessor worked independently using the ACJS for a few weeks to make the judgements it allocated to them. At the end of each round of judgements the system provided statistical and graphical output on such as the number of judgements and reliability coefficient. Some of this information was emailed to the assessors along with a summary explanation. When the overall reliability coefficient was determined to be high enough assessors were emailed to ask them to stop inputting judgements. Once again, assessors were supported by a set of instructions and access via email and phone to members of the research team.

Stage 4. Online meeting for review. A final synchronous online meeting was set up using the Adobe Connect video-conferencing system. The intention of this meeting was to provide a forum for presentation of the results from the ACJS, view a number of portfolios that had either scored high or low, or for which judgements appeared to be less reliable. There was also an opportunity for participants to report on their

experiences, and give their impressions of the ACJS and the pairwise comparison method.

Overall, almost all assessors were able to participate in all four stages, from home or school, and following the general instructions, with minimal personal support. As a result we were able to collect a range of data to analyse including the scores from Stages 3 and 4, researcher observations and anecdotal records, and interviews with the assessors upon completion. The findings are now discussed.

4. FINDINGS ABOUT ONLINE SOCIAL MODERATION

Findings from an analysis of the qualitative (interviews) and quantitative data (scores) are now reported in summary form starting with an analysis of the qualitative data. More detail on some of the data and analysis may be obtained from previous reports [12; 13]. This qualitative analysis included the interviews that focussed on attitudes and perceptions about the authenticity and quality of the digital representations, the ease and effectiveness of the comparative judgements process, and the online scoring for moderation and standard setting purposes. This analysis is followed by a discussion of findings from the quantitative data around the reliability of the scores, which is augmented with data from the notes assessors entered in ACJS while making their judgements, and a report from an expert assessor on a set of portfolios having less consistent judgements.

4.1 The fidelity of the digital representations

In the first two phases there had been a particular focus on perceptions of assessors, teachers and students of the authenticity and quality, or fidelity, of the digital representations. Because the work being digitised was artwork almost everyone would perceive that the digital representations could never be as good as viewing the original artwork. However, the question was whether the quality of the digital representations was adequate to reliably score the work for summative assessment purposes. In particular it was important that assessors had enough information to judge the work as they would when viewing the original work. On balance from the first two phases of the study it was determined that almost all the digital representations were more than adequate for assessment purposes. [Note: Our scores did not influence the final results.]

In general the evidence in the Phase 3 appeared to show that almost all of the assessors involved were able to visualise the original work by viewing the digital representations. For example, one assessor could readily identify where students 'understood the use of elements and principles of art'. From the interviews two were strongly of the opinion that the representations were more than adequate while one felt that for some artworks the representations were not adequate. The others held opinions between these two. However, to some extent they all held some concern that intricate features of the works such as 'textural nuances', 'size', 'techniques', and 'materials' may not be fully represented, although this may not affect the final judgement. Further, some of them felt that the quality of photographs, and in particular videos, could be improved with a view that the latter were really only useful to indicate the size of the artwork, particularly for 3D works. Despite the limitations they all supported the value of the pairwise comparison method with one assessor stating that this 'exhaustive method of comparative marking probably cancels out this problem as accuracy of marking seems evident'. So overall the consensus was that the digital representations had adequate fidelity for assessment purposes.

4.2 Pairwise judging process and ACJS

For almost all assessors in all three phases of the study the pairwise judging process was a new concept. They were familiar with using rubrics for analytical marking but not with the concept that data from a large number of binary judgements could be used to generate a score. Further, most of them had no experience in using online tools with either method or with using digital representations in these processes. However, they all relatively quickly developed an understanding of the mechanics of the pairwise judging process and that such relative judgements could be easier and more consistent than the absolute process using a rubric, particularly for the purposes of moderation. For example, one assessor made the following observation.

I found comparative [pairs judging] much easier than the analytical method. Because marking art can be subjective at times, having another piece to compare the work to allow the piece to be marked against something solid and 'real'.

They did recognise that comparisons were more difficult when the work was of similar quality and that it was sometimes difficult to make a judgement that balanced components of the holistic criterion. These components were readily represented in the analytical marking rubric but they had to be retained and balanced in the mind together for pairwise judging.

In addition to their perceptions of the concepts involved we were also interested in their experience of using the online assessment tools, particularly the ACJS. They all used these tools on computers either at their homes or schools. A few of them needed help from their school IT support, particularly to install the Firefox browser, and thus they used the tools from school. Although some complained about slow file downloads they all were able to use the tools and found them 'very easy and accessible', and 'easy to navigate'. Some did suggest that the ACJS could have a zoom function or full screen function for photographs and videos. Also they would have liked the pair of portfolios to be available continuously, side-by-side. However, overall we could conclude that it would be feasible to use pairwise judging with the ACJS for moderation purposes to involve teachers from across the state.

Overall it was concluded that the assessors preferred using the pairwise judging method in the manner it was facilitated by the ACJS. The system was relatively easy to use and they believed that they were able to more accurately and consistently judge the student work.

4.3 The value of the online meetings and support

The intention was to complete all moderation processes with no face-to-face meetings requiring the need for each stage to be facilitated through online systems and supported using online or phone communication. Assessors needed to be supported in using the two online assessment tools and joining the online meetings. The latter was particularly critical in developing an understanding for pairwise comparison, learning to use the ACJS and being able to consistently apply the holistic criterion. They all made some use of the instruction documents, and email or phone contact with the research team. Typically they found that the documents were 'referred back to' when needing to access the online systems. In addition some sought help from

their school 'IT department'. The final outcome was that all assessors were able to access the three online systems.

As may be expected the most difficulty was associated with using the Adobe Connect conferencing system for the online meetings because the Firefox browser was recommended, school firewalls had to be encountered and microphones were needed for audio conferencing. However, only one was not able to have an adequately functioning setup complaining that it was, 'very frustrating as we did not have the software to use and I was unfamiliar with the Connect conferencing site, the Firefox software and the process of having a video-conference'. While the majority used the systems from home because they felt the technology was more reliable and the environment was more conducive, some worked at school particularly if the Internet connection was better or there would be fewer interruptions.

In general the assessors, apart from two, found the initial online meeting to be very useful in providing opportunities to 'ask any questions directly relating to the process', showing 'how to use the software', and getting 'feedback'. At the time the researchers involved believed that the meeting had achieved the required outcomes, in particular all assessors were then able to use the ACJS. Most found the final online meeting 'good' probably because the meetings reduced the feeling of isolation associated with teaching in rural schools where they were often the only art teacher. Comments included that it was helpful to 'hear the input from other art teachers', a 'good way to have questions answered instantly' and 'good visuals to see how to make things happen'.

Overall we believe that we had demonstrated that social moderation could be adequately achieved without using face-to-face meetings.

4.4 Assessor perceptions of the moderation processes

For our approach to online social moderation to be implemented widely it would be necessary for teachers to perceive it to have adequate efficacy. Therefore, in interviewing participating teachers we asked about their perceptions of the processes and online systems for the purpose of moderation. In general they believed that either of the online scoring systems would be "an excellent way to moderate work" and "great for backing up decisions after in school and district moderation". To some extent this was probably due to the difficulty of rural teachers participating in social moderation, at one put it the current face-to-face moderation process was 'out-dated'. In fact some had not had previous opportunities to view artworks of students from other schools and thus the online tools were perceived to be "very effective" for standard setting purposes because assessors could see a "greater amount of work, viewed with the greater range, the better the understanding of standards".

As previously explained most perceived the pairwise comparison method as preferable for highly subjective areas such as art, with one stating that, "analytical moderation by itself is a waste of time but the comparative pairs marking could be very useful". The major concern of some was that using online tools meant that assessors were not seeing the original works that was perceived to be 'NOT the same at all'. However, in general almost all indicated that they perceived online social moderation in the way they had experienced it preferable to the status quo. To some extent this appeared to be not only the opportunity to participate but also that they perceived that the final results would be more reliable. One of them made the point that it was 'very reassuring that the marks given and comments made were similar to the ones I gave. It also gave me a wider view of the types of artworks being developed by students in the State which was helpful'.

For wider implementation this approach to social moderation needed to be demonstrated to be not only feasible but also economic. Therefore, they were asked for a record of the time taken for analytical marking, pairwise judging, and other assessment activities such as online meetings. The mean time they spent using the analytical marking system was 3.2 hours and the pairwise judgements system was 8.6 hours. They estimated that the time spent on online meetings and other activities took on average 3.2 hours. This is clearly more time spent than would be economically feasible although if more teachers were involved each would do far fewer pairwise judgements. Even so the results would have to be demonstrated to be clearly more reliable.

The assessors perceived that the moderation processes built around online tools were good for assessing visual arts student work. In particular it was a good way to involve those from disparate locations.

4.5 The reliability of the pairwise judgements

The purpose of moderation is to improve the reliability of scores or grades associated with an assessment. To investigate the reliability of scores generated by the pairwise judgement method statistical measures were used for each phase of the study. In addition in Phase 3 an expert assessor's qualitative judgements were also considered. The ACJS generated its own reliability statistics including a coefficient equivalent to a Cronbach's Alpha. In addition correlation analysis could be used in comparing the scores from the ACJS with those from analytical marking (within the study and the official external scores). Analyses in the first two phases of the study provided evidence that the pairwise judgement method generated reliable sets of scores for artworks. Because in Phase 3 the same portfolios were used as for Phase 1 (but different assessors) it is useful to initially consider the outcomes of this phase and compare these with those from Phase 3.

In the first phase the reliability coefficient from the ACJS was 0.96 and the scores generated correlated strongly with those from analytical marking and the official WACE marking ($r=0.80$ and 0.85, $p<0.01$). Interestingly the correlation between the scores from analytical marking by three assessors was poor (average $r=0.46$) although Rasch measurement analysis of the averaged scores yielded a Cronbach's Alpha coefficient of 0.94. A likely interpretation of this outcome is that the judgement of individual assessors is highly subjective in relation to the application of the criteria to specific artworks, however, their combined judgement is more reliable as represented by the analytical scores average or the pairwise comparisons judgements.

As for the others phases the intention for Phase 3 was to achieve a reliability coefficient from ACJS above 0.95, however, after 15 rounds it had only reached 0.88 and did not appear to be increasing. Therefore the process was stopped to allow analyses of all the data. Initially the scores were compared with those from Phase 1 yielding only a moderate correlation coefficient ($r=0.65$). Further, for some portfolios there were substantial differences between their rank position from the pairwise comparison judging in Phase 1 and Phase 3. Some of these differences in ranking can be explained by the fact that a small

change in score can lead to a large change in ranking, particularly if the range of scores is small. The range of scores in Phase 3 was 10, which was about 62% of the range in Phase 1 that was 16. It was decided to investigate the potential explanations for this discrepant outcome with additional analysis of the data including the notes that assessors typed into the ACJS as they recorded their judgements. These notes could be analysed by judge and by portfolio, and thus for a portfolio the notes of all assessors who had viewed the portfolio could be compared as an indication of their perception of that work.

We identified a small set of portfolios that showed a large difference in rankings between the two phases. The notes in ACJS indicated disparate views on the quality of the work with some seeming to focus more on art skills and others on artistic merit (i.e. the meaning of the work). For example for one portfolio an assessor typed "sound use of materials but that could have been pushed more" while another assessor viewing the same work typed "unique and creative, taking risks in design solutions". The conclusion was that artworks that were either only perceived to evoke meaning or demonstrate only high levels of skill were more likely to be inconsistently judged. This was not related to the type of artwork (e.g. 2D, 3D, painting). In addition we employed a highly experienced Visual Arts assessor to review this set of portfolios. She suggested that the scores generated by the Phase 1 assessors were more accurate and that the Phase 3 assessors demonstrated a lack of experience in assessing such work. Further, when we engaged an expert in Rasch measurement analysis to report on inconsistencies in judgements he concluded that this was more associated with particular assessors and from demographic data we had gathered it appeared that these assessors were those with the least experience in WACE marking (most had no experience). It appeared that the assessors in Phase 1 were more consistent because they were experienced WACE markers. The Phase 3 teachers were not as experienced and this showed in the quality and consistency of their judgements.

From this conclusion we formed the opinion that if we had have included one or two more online meetings during the judgement processes in ACJS to review particular judgements, then the quality of judgements would have improved and thus the final reliability. Thus the model for online social moderation we recommended includes these online meetings as shown in Figure 3 in steps (6), (8) and (9).

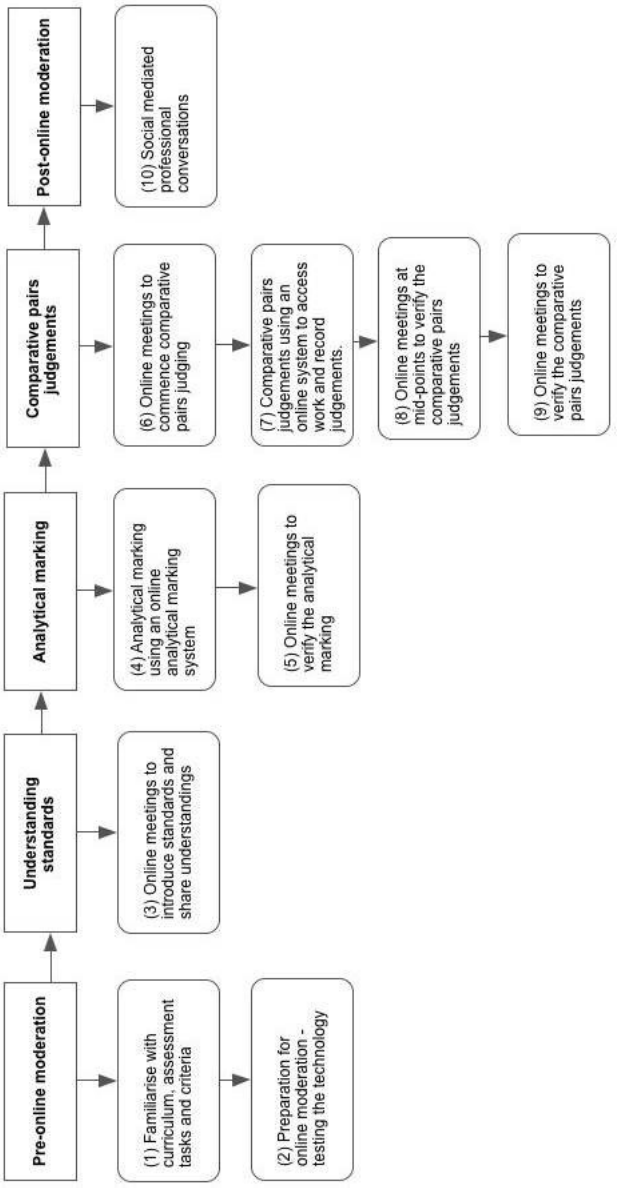

Figure 3: A model for online social moderation.

5. CONCLUSION

The findings of our study in terms of the use of online social moderation for the assessment of digital representations of artworks by senior secondary students are that technically it is feasible, but that the outcomes depend more on the experience and knowledge of the assessors. Typical teachers in Western Australia have adequate access to computers and the Internet to be able to use online scoring tools, access the digital representations and communicate using conferencing and other forms of electronic communications. As a result there would be no need for face-to-face meetings or teachers travelling long distances to view artworks. It was clear that the pairwise comparisons method of judging has advantages over analytical marking for highly subjective material such as artworks. However, the reliability of either method was dependent on the experience and knowledge of the assessors. Therefore the method we used would need to include more scaffolding through online meetings for more novice assessors.

Because there was evidence that such an approach provided teachers with the opportunity to develop their professional knowledge and understanding of standards and assessment criteria [2] we believe that with time the results would become highly reliable. How efficient this approach can be ultimately made will require further research into this model for online social moderation. In particular we aim to try variations on our model for online social moderation for other courses that have different types of practical assessment tasks and thus different forms of digital representations.

6. REFERENCES

[1] Adie, L.E., 2013. The development of shared understandings of assessment policy: Travelling between global and local contexts. *Journal of Education Policy 29*, 4, 1-14.

[2] Adie, L.E., Klenowski, V., and Wyatt-Smith, C., 2012. Towards an understanding of teacher judgement in the context of social moderation. *educational Review 64*, 2, 223-240.

[3] Andrich, D., 1988. *Rasch models for measurement.* Sage Publications, Newbury Park.

[4] Filemaker Inc., 2007. Filemaker Pro 9 Filemaker, Inc., Santa Clara, CA.

[5] Harlen, W., 2007. *Assessment of learning.* Sage Publications., London.

[6] Heldsinger, S. and Humphry, S., 2010. Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher 37*, 2, 1-20.

[7] Hipkins, R. and Robertson, S., 2012. The complexities of moderating student writing in a community of practice. *Assessment Matters 4*, 30-52.

[8] Humphry, S.M., Wray, W.H., and Wray, F.W., 2013-2015. Pair-Wise Web Software. The University of Western Australia., Perth, Western Australia.

[9] Kimbell, R., 2012. Evolving project e-scape for national assessment. *International Journal of Technology and Design Education 22*, 2, 135-155. DOI= http://dx.doi.org/10.1007/s10798-011-9190-4.

[10] Klenowski, V. and Wyatt-Smith, C., 2010. Standards-driven reform years 1-10: Moderation an optional extra? *The Australian Educational Researcher 37*, 2 21-39.

[11] Malone, L., Long, K., and De Lucchi, L., 2004. All things in moderation. *Science and Children 41*, 5, 30-34.

[12] Newhouse, C.P., 2014. Using digital representations of practical production work for summative assessment. *Assessment in Education: principles, policy & practice 21*, 2, 205-220. DOI= http://dx.doi.org/10.1080/0969594X.2013.868341.

[13] Newhouse, C.P. and Tarricone, P., 2014. Digitizing practical production work for high-stakes assessments. *Canadian Journal of Learning and Technology 40*, 2, 1-17.

[14] Pollitt, A., 2012. The method of adaptive comparative judgement. *Assessment in Education: principles, policy & practice 19*, 3, 281-300.

[15] Rasch, G., 1961. On general laws and the meaning of measurement in psychology. In *The fourth Berkeley symposium on mathematical statistics and probability*, J. NEYMAN Ed. University of California Press, Berkeley, California, 321-333.

[16] Smith, C., 2012. Why should we bother with assessment moderation? *Nurse Education Today 32*, 45-48.

[17] Van Der Schaaf, M., Baartman, L., and Prins, F., 2012. Exploring the role of assessment criteria during teachers' collaborative judgement processes of students' portfolios. *Assessment & Evaluation in Higher Education 37*, 7, 847-860.

[18] Wilson, M., 2004. Assessment, accountability and the classroom: A community of judgement. In *Towards Coherence between Classroom Assessment and Accountability*, M. WILSON Ed. University of Chicago Press, Chicago, IL, 1–19.