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).

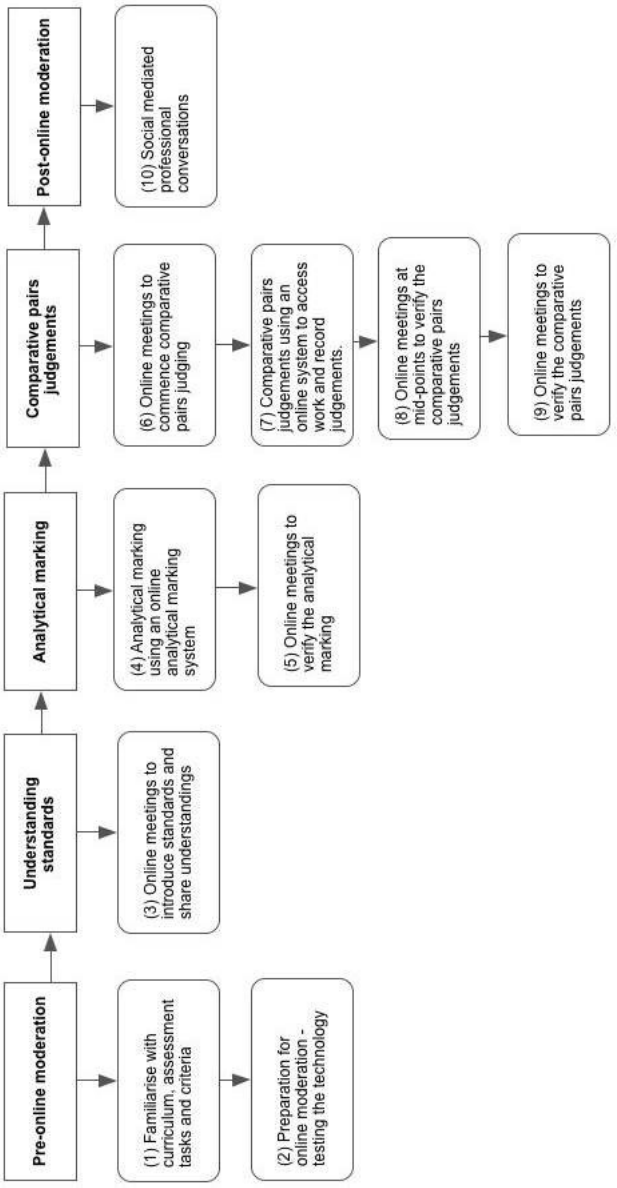

Figure 3: A model for online social moderation.

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.