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).

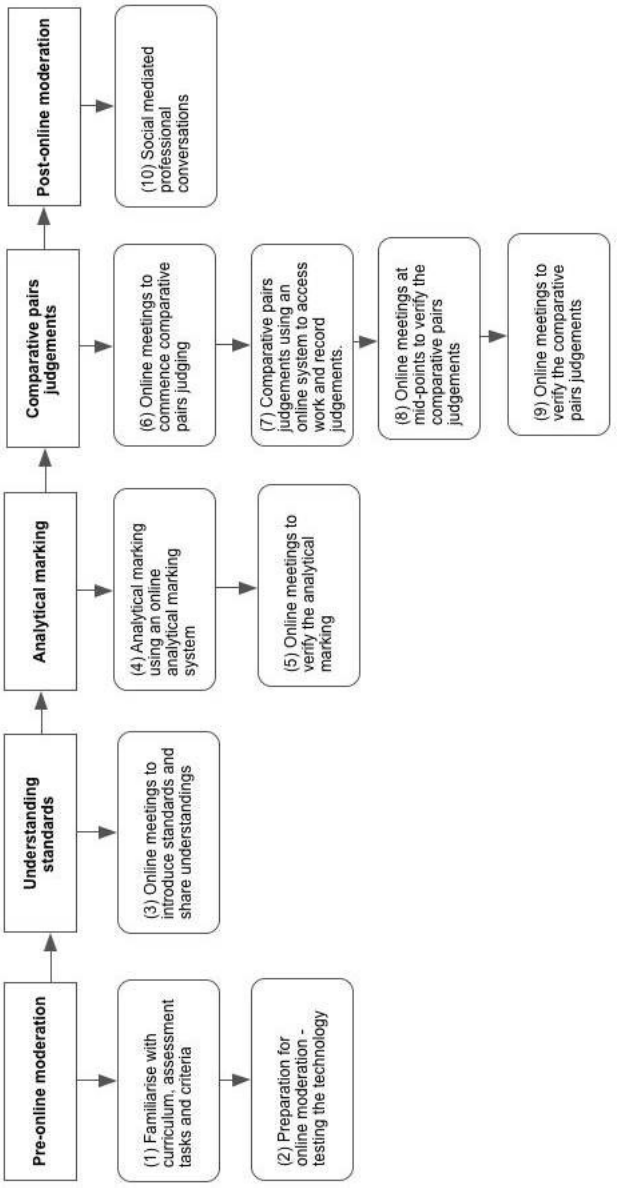

Figure 3: A model for online social moderation.

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.