2016

Online moderation of external assessment using pairwise judgements

Christopher Paul Newhouse
Edith Cowan University, p.newhouse@ecu.edu.au

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

Online Moderation of External Assessment using Pairwise Judgements

C. Paul Newhouse & Pina Tarricone
School of Education, Edith Cowan University

Abstract: When an assessment involves judgement by more than one assessor, it is usual to consider that some form of moderation is required to ensure consistent results. This often involves face-to-face meetings to compare judgements, and agree upon a score or grade. As well as inconsistency between assessors there is typically substantial error due to inconsistency in judgement between student performances by individual assessors. This is particularly the case where judgements are necessarily highly subjective. The traditional forms of moderation tend to be logistically difficult and analytical scoring using devices such as rubrics make generating reliable scores difficult. It is likely that problems such as these can be tackled using modern technologies, a thesis that we set out to investigate in the final phase of a three-year project into summative assessment in senior secondary schooling. The study investigated the use of digital technologies to support a form of social online moderation, which involved the use of digital communications and assessment tools to facilitate a pairwise comparative judgement approach. It involved a small group of Visual Arts teachers from rural schools in Western Australia assessing digitised forms of artworks submitted for high-stakes summative assessment at the end of Year 12. The aim was to determine whether the use of these technologies would provide good moderation outcomes and valuable professional learning for those involved. The participants were guided through the processes that involved no face-to-face meetings and were questioned about their experiences of these processes and technologies, and their attitudes and perceptions of them. In addition, the scores from their judgements were analysed. The results demonstrated that it was feasible to use these technologies to support moderation processes under these circumstances. Further participants perceived that the social online moderation processes supported by the technologies had assisted them in making more consistent judgements and increasing their understanding of the standards of work submitted and the application of the assessment criteria. They found the comparative judgement approach easy to use and appropriate for assessing the highly subjective artworks. However, the study found that probably due to the inexperience of some of the assessors the reliability of the final scores was not as high as anticipated. As a result, a slightly more rigorous method for social online moderation is recommended by the study.

For eight years our research centre, Centre for Schooling and Learning Technologies (CSaLT), investigated the use of digital technologies to support improvements in summative assessment in senior secondary schooling. We have worked on this with our industry partner the School Curriculum and Standards Authority (SCSA) of Western Australia. In particular, the aim was to improve the authenticity of assessment in terms of what is assessed, and the validity and reliability of the assessment in terms of how the performance of a student is judged to generate a score. In addressing the latter, we investigated the use of pairwise comparative judgements supported by online systems, and found that this approach can be easier for assessors, and generate more reliable scores, particularly where the performance is highly subjective. An outcome of this was that assessors reported that they perceived that the approach added to their professional learning through increasing their awareness of the standards of performance, and understanding of the application of assessment criteria. Therefore, we decided to add to our investigation the application of our approach to moderation processes. This was particularly motivated by the decision of SCSA to introduce 'general' non-ATAR versions of senior secondary courses that would have school-administered assessment, moderated by a common 'externally set task' for each course.

We saw an opportunity for teacher-based online moderation of these 'externally set tasks' using pairwise comparative judgements that would allow a large variety of types of tasks (i.e. not just paper-based tests as was intended). In the past for such situations moderation had involved face-to-face meetings of teachers to check their marking of the tasks and ensure that standards were maintained across the vast expanses of Western Australia. Researchers in our centre perceived that this approach would be problematic and that the approach to assessment that we had investigated using online systems and digitised representations of student work may be better. Therefore, in the final phase of our research we decided to investigate the potential of digital technologies, or ICT (Information and Communications Technology), to facilitate social online moderation to generate reliable scores and provide professional learning for assessors. In particular, due to the expanse of Western Australia we wanted to demonstrate that this could be achieved no matter where the assessors resided. In this paper we describe an investigation of the assessment of digitised artworks, using both analytical and pairwise comparative judgements methods of scoring, in a social online moderation context.

What is social online moderation?

When an assessment involves judgement by more than one assessor, it is usual to consider that some form of moderation is required to ensure consistent results. This often involves face-to-face meetings to compare judgements, and agree upon a score or grade. As well as inconsistency between assessors there is typically substantial error due to inconsistency in judgement between student performances by individual assessors. This is particularly the case where judgements are necessarily highly subjective. Traditionally in countries such as Australia moderation has relied on a form of social moderation, involving physical meetings between small groups of teachers to review samples of student work, and gain a consensus understanding of achievement standards (Malone, Long, & De Lucchi, 2004). In addition to improving the reliability of scores these social moderation meetings are also important professional learning opportunities for teachers to review their professional knowledge and understandings, and share expertise in instruction and assessment (Adie, Klenowski, & Wyatt-Smith, 2012). However, the benefits of the traditional moderation meetings are limited to the time of the meeting, with little further professional discussion possible. It is therefore likely that online technologies could be used to improve the outcomes of social moderation where a ‘community of practice’ is able to share expertise, develop understandings of the assessments and achievement standards (Adie, Klenowski, & Wyatt-Smith, 2012; Hipkins & Robertson, 2012; Smith, 2012; Wilson, 2004). This social online moderation may use both asynchronous (e.g. email) and synchronous (e.g. video-conferencing) online systems allowing teachers to be engaged with the processes and the community of judgement, no matter where they are located (Adie, Klenowski, & Wyatt-Smith, 2012; Wilson, 2004).

The main aim of social online moderation is the same as for face-to-face moderation, to ensure consistency in the understanding and assessment of student work. However, the former provides unique opportunities to exemplify specific qualities in student work through digital representation, whilst engaging teachers, in all locations, in communication and assessment activities. Although the benefits of social online moderation have been discussed it has not been fully adopted by any educational system (Adie, 2013; Adie, Klenowski, & Wyatt-Smith, 2012). However, there are a few studies that have investigated the use of social online moderation, for example, Adie, Klenowski and Wyatt-Smith (2012) involved a group of 50 Queensland teachers from 21 rural schools, in social online moderation. They used the Internet and a telephone to participate in online moderation meetings, facilitated by the Cisco WebEx video-conferencing system. The study found that such social online moderation could increase reliability and the consistency of understanding of standards, more broadly than their own school or district context. A separate study by Klenowski and Wyatt-Smith (2010) investigated the use of social online moderation to improve the consistency of judgements made by teachers and their understanding of assessment standards. In turn this helps them to make adjustments to teaching strategies allows them to better inform their students, parents and the wider community.

Our study used the social online moderation context with the use of the pairwise comparative judgement method of scoring with the aim of assisting teachers to develop a shared knowledge and understanding of the standards. The method of comparative judgement (pairwise comparison) is based on Thurstone’s (1931) law but it has only been practical to use for large assessment samples with the development over the last decade of software systems such as the Adaptive Comparative Judgement System (ACJS) (Pollitt, 2012) and the Pair-Wise Web Software (Humphry, Wray, & Wray, 2013-2015). However, there has been little use of these systems to support social online moderation. Comparison is fundamental to all measurement, including educational assessment where this can be between two performances or between a performance and a theoretical standard (e.g. a set of marking criteria). Our study used the ACJS that generates pairs of digital portfolios for each assessor to judge, and provides a function for judgements to be made concerning which portfolio is the better of the two, and an area for assessors to record their own private notes about each portfolio. The system is adaptive, meaning that the pairs are generated dynamically based on the results of previous judgements, and it calculates the scores and the reliability coefficients after each ‘round’ using Rasch modelling algorithms.

Method for study

Researchers from 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 conducted a three-year study into the use of digitised portfolios of creative work for summative assessment in the WA Certificate of Education (WACE) courses of Design, and Visual Arts. This was supported by an Australian Research Council (ARC) Linkage research grant. The study was conducted in three phases and involved the assessment of digitised student production work as portfolios, including artefacts (e.g. paintings, sculpture, drawings, and photographs), using both an analytical and a comparative judgement method of scoring. The research design and results of the first two phases of the study are reported in a number of publications (Newhouse, 2014; Newhouse & Tarricone, 2014). These phases were designed to firstly test whether the student work for assessment in the two courses could be represented adequately in digital form and scored online using the two methods, and secondly to test whether students could digitise their own work and submit it online.

This paper is concerned with the third phase of the study that aimed to investigate the effectiveness of using online communication systems and comparative judgement systems with digitised Visual Arts portfolios for the purpose of moderation and professional learning of standards. The sample for the phase comprised 12 Visual Art teachers, with a range of experience, from rural schools. They work to be assessed was 75 digitised Visual Arts submissions that had resulted from the first phase of the study. These had been submitted as physical artworks with required documentation to SCSA for examination and then had been digitised into digital files by the research team. These files were stored on a server for analytical marking using a custom-built online tool, and uploaded into the ACJS for comparative judgement (Newhouse, 2014). The aim was to support these teachers to use these online technologies to complete a social moderation method over a period of weeks either from their schools or homes.

The plan for social moderation of the Visual Arts digital portfolios followed the following sequence.

1. Each teacher independently used an online analytical marking method with a sample of 10 portfolios and a custom built tool in the Filemaker Pro database system. This was designed to familiarise them with the required assessment criteria and the rank of quality of work because the sample had been selected to represent this range, as determined by the ranking from scores in the first phase of the study.
2. All teachers joined a synchronous online meeting supported by the Adobe Connect video-conferencing system. At this meeting the 10 sample portfolios were reviewed, they were introduced to the concept of comparative judgement and the operation of the ACJS, and as a group made judgements of the first few pairs of portfolios. This allowed them to discuss the basis on which they would make a judgement of the winning portfolio. The judging criterion was developed based on the criteria from the WACE practical submission used in the analytical marking.
3. Each teacher independently used the ACJS over a number of weeks to enter the judgements allocated to them by the system. All 75 portfolios were involved in this process and at the end of each round of judgements all teachers were emailed a summary of progress (total number of judgements made by each teacher, numerical and graphical representation of reliability of scores).
4. All teachers joined a synchronous online meeting supported by the Adobe Connect video-conferencing system. At this meeting the final results from the ACJS were presented and discussed, illustrative portfolios were reviewed, and there was an open discussion on the relative merits of the approach and the performance of the technologies used through the sequence.

Throughout this sequence the teachers were supported by digital documents and individual email or telephone assistance. In general, it was found that with a prior practice run all teachers were able to participate with the audio-visual conferencing system for the meetings and then, with little difficulty, follow the instructions to use the two online scoring systems.

A range of data was collected including researcher observations; interviews with the teachers upon completion; and the scores generated by the two methods of scoring. Qualitative and quantitative analyses were conducted using these data. Details of the results of the analysis from the quantitative data (i.e. scores) have been reported (Tarricone & Newhouse, 2016). However, a summary is provided at the end of the section reporting results in a discussion of the reliability of the scores. This is preceded by a discussion of the results from an analysis of the qualitative data, in particular the interviews with teachers. These interviews were conducted to elicit attitudes and perceptions about the authenticity and quality of the digital representations, the ease and effectiveness of the comparative judgements process, and on online scoring for moderation and standard setting purposes. In addition, notes they entered into the ACJS while making judgements, and a report by an expert assessor were used to provide information about the portfolios that were identified as having inconsistent judgements.

Findings concerning social online moderation

Here an analysis of the qualitative data is presented firstly, in particular from the final interviews with the teachers, and then a summary of an analysis of the quantitative data in relation to the reliability of the scores.

The results are presented in terms of: the authenticity and quality of the digital representations; the comparative judgements process and online tool; the quality of student work; the support for online scoring; and the reliability of pairwise comparisons.

Digital representations: authenticity and quality

Overall the teachers considered that there was a good range of artworks from ‘poor to very strong’ in line with expectations. One explained that the students showed that they ‘understood the use of elements and principles of art, some used them in simple ways, others in complex ways -- ranging from unimaginative in terms of innovation and communication through to excellent’. They identified a number of factors which influenced their judgement of the standard of the artwork including: creativity, technique, refinement of the work; originality and evidence of development of final ideas; quality of presentation; ability to portray the desired meaning of the artwork, supported by the artist statement; experience teaching, marking and moderating work; not seeing the full range of work to be judged at one time; quality of photos and video; and quality and articulation in the artist statements.

The teachers who completed the comparative judgements generally reported that the quality of the digital representations was adequate and fair. However, some reported that some of the images did not represent the details, materials, textures and dimensions of the artworks. Further, they reported that the videos were generally out of focus, unclear, too small, and only really indicated the size of the artwork. Also they felt that the objects behind the works distracted viewing the artwork both for the digital images and the video. Two were adamant that the digitised artworks were clear enough to ‘assess the works well’, while for one ‘most were fine’ and ‘some were not’.

In general they were concerned that intricate features of the works such as ‘textural nuances’, ‘size’, ‘techniques’, and ‘materials’ may not be fully represented. Although they felt that the digital representations had some limitations, including the loss of the real-life impact of the artwork, they were in support of the comparative judgements approach, for example, stating that the ‘exhaustive method of comparative marking probably cancels out this problem as accuracy of marking seems evident’.

Pairwise judging process and online tool

In general the teachers found the comparative judging tool to be ‘very easy and accessible’, and ‘easy to navigate’, but some were frustrated with slow download of files whether at home or school. A small number required help from either their school’s IT support or from our team, with the initial setup, in particular installing the Firefox browser that was preferred for using the system. They made a few suggestions to improve the functionality of the ACJS online tool. For example, it was suggested that the viewer window for photographs could have a zoom function and videos could be ‘seen at full screen size’. The most common suggestion was that the system could ‘have the works compared on the one screen – side by side’.

For all of them the pairwise comparative judgement process was new, however, they quickly developed an understanding for the process. By the end they all made comments about the judgement process using terms such as ‘super easy’, ‘fine’ and ‘a simple and clear process’. They were able to compare it with the standard analytical marking process that they were familiar with. Overall they all preferred the pairwise comparative process, particularly when the purpose was for moderation. The following are indicative of their views.

> *Comparative [pairs marking] done in great numbers, as we did, seems to weed out inaccuracies better. The comparative seems to have less of personal preferences having influence over marking. Comparative marking helps sort work into a range of marks more accurately. But maybe statistics indicate this better.*
>
> *I feel the comparative pairs judging is better [than analytical marking]. When marking in the classroom/moderating with other teachers I always compare artworks. It gives a better judgement of where each artwork is placed in comparison to the others. I generally line them up in order of marks allocated. The comparative pairs judgement was the closest to this.*
>
> *I believe comparative [rather than analytical] because it puts me in the same position as if I was marking work in class and had to put the work presented on a scale. I think analytical is good because it has solid structure in its marking process, and that comparative marking is a more reliable system for marking.*
>
> *I found comparative [pairs marking] 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'.*

Some made statements that showed that they recognised that the process was more difficult 'when comparing the work of two students who were quite similar'. Further, some made suggestions to improve the judging process such as having the criteria and assessment task on-screen (they did have a separate document). There were also suggestions about the digitised materials they were judging such as to improve the quality of some photographs and videos. There were mixed views of the value of the videos with one suggesting that it was of no value on 2D artwork and somewhat helpful on 3D artwork. Fortunately only one complained that technology functionality was a problem, stating 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'. In summary, given that for all of them this was a new method of scoring and that none of them had used the main online systems before, what we had attempted was clearly feasible as a replacement for the more traditional face-to-face analytical marking moderation paradigm.

Online meetings and other forms of support

The majority connected to the online meeting from home for a number of reasons, including: better technology reliability, computer and Internet, at home compared to at school; a quiet working environment at home; and time constraints and interruptions at work are problems. Those that connected from school explained that: there were less interruptions at school; the technology, including the Internet connection, was better at school than at home; and they had access to all the required materials and software at school than at home.

Nearly all of the teachers commented that the initial online meeting was helpful as it was visual and that it clearly demonstrated the 'marking' process and provided an opportunity 'to ask any questions directly relating to the process'. Further this meeting 'answered a lot of questions related to the marking process', 'showed how to use the software', 'was able to test and get feedback', 'demonstrated the process for marking clearly' and 'provided the opportunity to ask any questions directly relating to the process'. One teacher stated that the 'last online session was good'. However, two found the meetings somewhat unnecessary. In generally the online meetings were perceived to help to reduce the feeling of being isolated in a country school as 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'.

They were also supported with instruction documents, and email or phone contact with the research team. They all indicated that these documents were helpful, clear and needed. For example, one stated the documents were 'always referred back to' when she 'had forgotten how to access' the online systems. Similarly they all stated that the email support was necessary for 'any questions that could arise during marking'. Some also referred to support they received from their school 'IT department'. The result from all forms of support was that they were all able to access the systems from workplaces and/or homes to enter their judgements.

They were asked how many hours they took 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 comparative judgements system was 8.6 hours. They estimated that the time spent on online meetings and other activities took on average 3.2 hours.

Perceptions of efficacy for moderation and standard setting

The participating teachers were asked for their perceptions of the efficacy of the processes and systems used for moderation and standard setting. In general they indicated that felt that the use of the online scoring systems (analytical and pairwise comparisons) 'would be an excellent way to moderate work' and 'great for backing up decisions after in school and district moderation'. One stated that the current moderation process was 'out-dated' and was difficult for rural teachers because of the travel required and limited time provided. All tended to echo these concerns as rural teachers. Another stated that the online scoring provided the opportunity to view and assess more artworks than the current moderation processes allowed. However, one highlighted a concern regarding not seeing the original works, stating that 'It's NOT the same at all'. Another expressed a preference for comparative judgement stating that: 'I think that the analytical moderation by itself is a waste of time but the comparative pairs marking could be very useful'.

They all indicated that using the online scoring systems would be 'very effective' for standard setting purposes. It would help assessors see a 'greater amount of work, viewed with the greater range, the better the understanding of standards'. One stated 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'. Other comments included the following.

> *Very effective because it gives art teachers a bank of information regarding the standard of work being produced in Western Australia and allows us to develop a better online community which would greatly help remote schools.*
>
> *As a recent graduate, working in the country, I found the experience extremely valuable. I was able to see the range of standards from other schools and make a judgement where my students sit.*
>
> *Being involved gave me a better overview of what is happening in this State as I am unable to attend the annual year 12 marking in the city'*

In general they clearly supported the use of online technologies to facilitate shared understandings, professional discussions and learning. Individuals also commented that they 'really enjoyed the process', 'it was great and very insightful' and the 'way in which it was run was easy to access and time given was appropriate to a busy teacher's life'.

Reliability of scores from comparative judgements

To be useful for moderation purposes not only do the systems need to be easy for teachers to understand and use but the resulting scores need to be adequately reliable. Overall in the first two phases of the study, it was demonstrated that a reliable set of scores could be achieved using comparative judgement for subjective materials such as Visual Arts portfolios. In the first phase the reliability coefficient (equivalent to Cronbach's Alpha) 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, $p<0.01$). However, in the third phase the reliability coefficient from the ACJS stabilised at 0.88 and was not climbing in subsequent rounds so the process was stopped after 15 rounds. Additionally, there was only a moderate (but statistically significant) correlation between the scores generated by pairwise judgements in Phase 1 and those in Phase 3 of the same Visual Arts portfolios (by different assessors). Therefore we investigated the possible reasons for this outcome by further analysis of all the data starting with the notes that the teachers had typed into the ACJS during the judgement of each pair of portfolios on the qualities in each portfolio and the basis for their decision on which was the better. These notes could be analysed by judge and by portfolio, with the latter allowing a comparison between the notes of teachers who had viewed the same portfolio.

We initially identified a small set of portfolios that showed a large difference in rankings based on scores from the ACJS between Phase 1 and Phase 3. Generally from the notes of assessors in the system there were disparate views on the quality of the work. A potential explanation for the differences appeared to be that some assessors had tended to focus on one of either skills or artistic merit (meaning of the work), rather than a balance of both. For example, for one of these submissions while one assessor noted a 'sound use of materials but that could have been pushed more' another noted 'unique and creative, taking risks in design solutions'. It seemed that artworks that could evoke significant meaning but that may be perceived to require low levels of effort or skill, and vice-a-versa, were more likely to be inconsistently judged. Further investigation of the type of artwork (e.g. 2D, 3D, painting, etc.) and what proportion of the assessors judged the work in various rounds (particularly the early rounds) did not lead to any conclusive findings.

Finally, an expert assessor (experienced in assessing tertiary entrance examination) was employed to view the identified set of portfolios and suggest reasons why assessors may disagree on the quality of the work. In general she suggested that the Phase 1 assessors judgements were more accurate and the Phase 3 teachers demonstrated a lack of experience in assessing such work. Taking all these analyses into account it was decided that the most likely reason for the reliability coefficient not improving further, and the lower than expected correlation between Phase 1 and Phase 3 scores, was the lack of a consensus understanding of the assessment criteria among some of the teachers in Phase 3. Also, it was noted that those that were identified as 'misfits' by the ACJS statistical analysis tended to be the less experienced with WACE marking. It appeared that the assessors in Phase 1 were more consistent because they were experienced WACE markers and the Phase 3 teachers were not as experienced and this showed in the quality and consistency of the judgements.

As a result of this finding we believe that these teachers needed more online meetings, both before and during the judgement processes. Although the meeting in which the judging processes and ACJS were introduced was perceived to be successful, it was not adequate for a consistent application of the assessment criteria. It is likely that if two or three additional meetings had been convened early in the use of the ACJS to review particular judgements then the quality of judgements would have improved and thus the reliability coefficient. Thus the model for social online moderation that we recommend from our study adds these online meetings to our implemented method. A schematic diagram of this model is show in Figure 1. The additional meetings are shown in steps (6), (8) and (9).

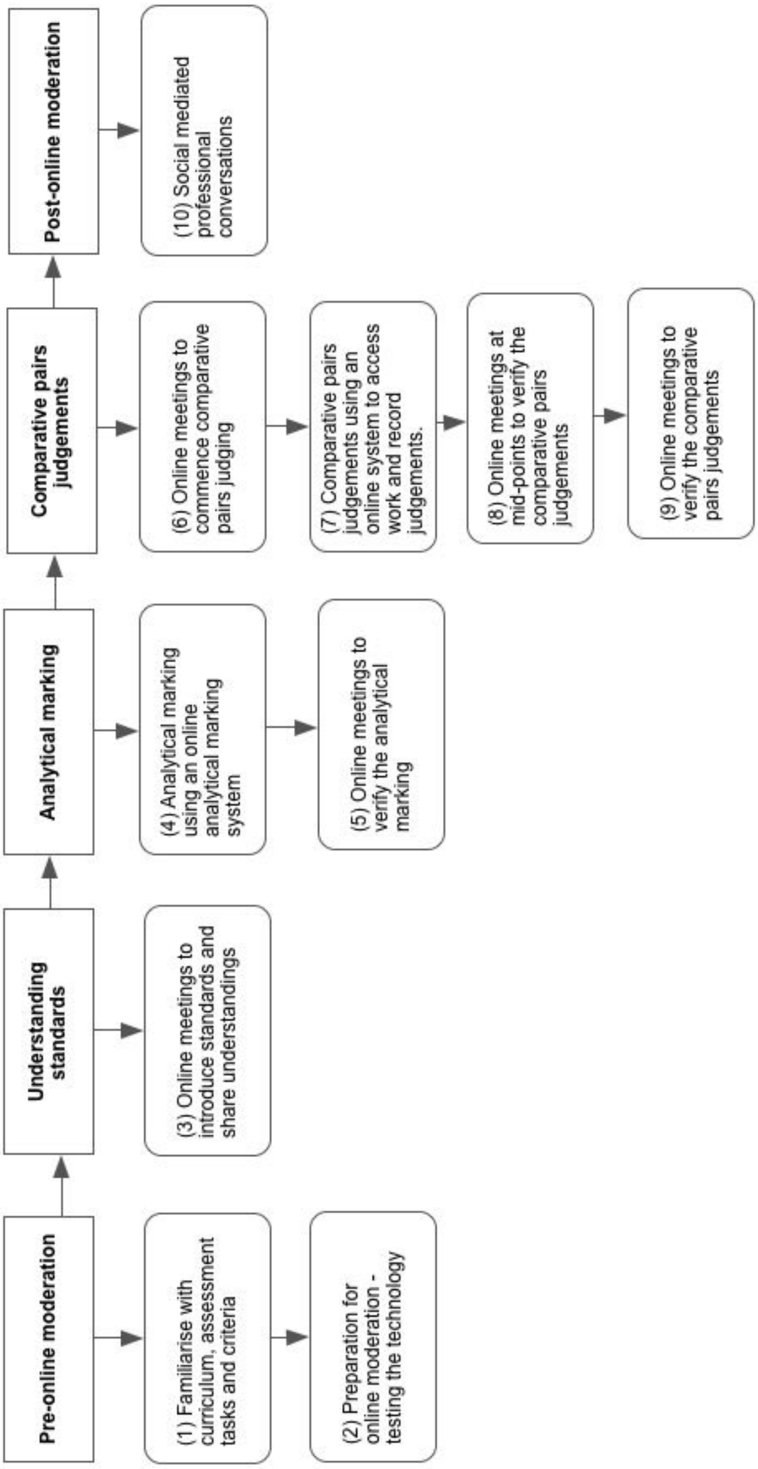
Pre-online moderation
(1) Familiarise with curriculum, assessment tasks and criteria
(2) Preparation for online moderation - testing the technology
Understanding standards
(3) Online meetings to introduce standards and share understandings
Analytical marking
(4) Analytical marking using an online analytical marking system
(5) Online meetings to verify the analytical marking
Comparative pairs judgements
(6) Online meetings to commence comparative pairs judging
(7) Comparative pairs judgements using an online system to access work and record judgements.
(8) Online meetings at mid-points to verify the comparative pairs judgements
(9) Online meetings to verify the comparative pairs judgements
Post-online moderation
(10) Social mediated professional conversations

Figure 1: Schematic representing a model for social online moderation.

Conclusion

The first two phases of our study considered the feasibility of using online scoring tools and digital representations of student work for assessment in the Design and Visual Arts senior secondary courses in WA.

The third and final phase investigated the potential of social online moderation, particularly as supported by digital scoring tools and the pairwise comparative judgement method. It was envisaged that in addition to the potential to replace traditional face-to-face approaches to moderation this would provide teachers with the opportunity to develop their professional knowledge and understanding of the quality of the student created artworks as envisaged by Adie et al (2012). Our study demonstrated that digital scoring tools, video-conferencing and other forms of online communication could be used with teachers in disparate locations for the purposes of social online moderation. Teachers were able to use the systems and perceived that the approach was successful for their own professional learning and for the purposes of moderation. In particular they believed that the pairwise comparative judgement approach was most appropriate for assessing artworks. However, the scores generated were not as reliable as anticipated with the likely reason being the relative inexperience of many of the teachers involved. This led to a revised model for social online moderation that incorporated the use of more online meetings during the judging processes.

In conclusion, we suggest that social online moderation, supported by the pairwise comparative judgement method, has potential for use in high-stakes summative assessment, particularly in practical courses such as Visual Arts where judgements are likely to be highly subjective. Social online moderation could replace other traditional forms of moderation to support teachers in rural schools and to develop a community of judgement with teachers across a region or state. The use of pairwise comparative judgements can help develop teachers' assessment and judgement skills, increase the reliability of judgements, validate teacher practice, and help improve teaching practice (Adie, 2011). However, it is clear that the critical factor in determining success is not the source material being assessed, but the judges sharing a common understanding of the assessment criteria. Therefore it is important now that further research is conducted in this area to test the validity of the model for social online moderation we have recommended from our study.

References

Adie, L. E. (2011). An investigation into online moderation. *Assessment Matters, 3*, 5-27.

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.

Adie, L. E., Klenowski, V., & Wyatt-Smith, C. (2012). Towards an understanding of teacher judgement in the context of social moderation. *Educational Review, 64*(2), 223-240.

Hipkins, R., & Robertson, S. (2012). The complexities of moderating student writing in a community of practice. *Assessment Matters, 4*, 30-52.

Humphry, S. M., Wray, W. H., & Wray, F. W. (2013-2015). Pair-Wise Web Software. Perth, Western Australia: The University of Western Australia.

Malone, L., Long, K., & De Lucchi, L. (2004). All things in moderation. *Science and Children, 41*(5), 30-34.

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: 10.1080/0969594X.2013.868341

Newhouse, C. P., & Tarricone, P. (2014). Digitizing practical production work for high-stakes assessments. *Canadian Journal of Learning and Technology, 40*(2).

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

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

Tarricone, P., & Newhouse, C. P. (2016). Using comparative judgement and online technologies in the assessment and measurement of creative performance and capability. *International Journal of Educational Technology in Higher Education, 13*(1), 1-11. doi: 10.1186/s41239-016-0018-x

Thurstone, L. L. (1931). The measurement of social attitudes. *The Journal of Abnormal and Social Psychology, 26*(3), 249-269.

Wilson, M. (2004). Assessment, accountability and the classroom: A community of judgement. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (pp. 1–19). Chicago, IL: University of Chicago Press.