progressions, making it hard to tell, for example, whether a poor performance in a short, formative test is a random aberration, or evidence of a fundamental misunderstanding and a real learning need. Developing tools of analysis and communication that can deal with this inevitable ambiguity is tricky. We should investigate the validity of machine-learning outputs, and whether they are aligned with alternative sources of evidence. And, we must evaluate the impact of data-fuelled approaches and machine learning products as they are introduced – and look for unintended consequences.

Cambridge Assessment has long been data driven. Big data, the convergence of teaching, learning and assessment, and the increasingly sophisticated operationalisation of machine learning and of data science more generally, are creating real opportunities for improving our understanding and practice of education. We should never put our faith in black boxes, however, nor introduce wide-scale change without evaluation. We must earn public trust by establishing and upholding clear ethical principles in relation to our use of data; be open; communicate continuously about what we are doing and why; inspire people with our vision and respond to their concerns; and always remember that we rely on their consent.

References

Benton, T. (2017). The clue in the dot in the 'i': Experiments in quick methods for verifying identity via handwriting. *Research Matters: A Cambridge Assessment publication, 23*, 10–16.

Cheung, K., Xu, J., & Lim, G. (2017). *Linguaskill: Writing Trial report*. Retrieved from https://www.cambridgeenglish.org/Images/466042-linguaskill-writing-trial-report.pdf

The Guardian (2019). *The Cambridge Analytica Files*. Retrieved February 8, 2019 from https://www.theguardian.com/news/series/cambridge-analytica-files

Herold, B. (2016, January 11). The Future of Big Data and Analytics in K-12 Education. *Education Week, 35*(17). Retrieved from https://www.edweek.org/ew/articles/2016/01/13/the-future-of-big-data-and-analytics.html

Lee, D. (2016, March 26). Tay: Microsoft issues apology over racist chatbot fiasco. *BBC News*. Retrieved from https://www.bbc.co.uk/news/technology-35902104

Singer, N. (2014, April 21). InBloom Student Data Repository to Close. *The New York Times*. Retrieved from https://bits.blogs.nytimes.com/2014/04/21/inbloom-student-data-repository-to-close/

Moderating artwork: Investigating judgements and cognitive processes

Lucy Chambers, Joanna Williamson Research Division and **Simon Child** Cambridge Assessment Network

(The study was completed when the third author was based in the Research Division at Cambridge Assessment)

Introduction

For the majority of standardised summative assessments in the UK, candidates will sit examinations. However, for certain practical or performance-based components, candidates will complete a non-exam assessment, which is marked by their teachers. To ensure that the standards of marking are the same across centres[1], samples of candidates' work from each centre are externally moderated. This process entails moderators, appointed and trained by awarding organisations, viewing the work and deciding whether the teachers have marked accurately and consistently. The aim of this study was to explore the cognitive processes and resources used by moderators when making judgements about artwork submitted for moderation.

The moderation method used by awarding organisations in the UK is that of inspection (see Joint Council for Qualifications[2], 2018, for a description of the moderation process). When making their judgements, moderators must consider the sample in the context of the centre as a whole, looking for trends and patterns in the marking. The moderators can make adjustments to the centre's marking, if necessary, to maintain the same marking standard across all centres. This must not be done with a view to changing the marks of individual candidates in isolation, but with a view to ensuring that the agreed standard is applied to all candidates (see Gill, 2015) for details of how centre-level mark adjustments are made).

Few studies have explicitly examined the cognitive processes involved in moderation. The only such studies that we are aware of are those of Crisp (2017) and Cuff (2017). The components under consideration in these studies involved the submission of mostly written work. The aim of this study was to investigate whether their findings hold when moderating submitted work of a very different nature, namely for Art and Design. There is little research on the marking and moderation of artwork. In fact, reviews observe that there is little detailed or technical research on assessment in art altogether (Gruber & Hobbs, 2002; Haanstra, Damen, Groenendijk, & van Boxtel, 2015; Herpin, Washington, & Li, 2011; Mason, Steers, Bedford, & McCabe, 2005).

Subject-specific research is particularly necessary for assessment in Art and Design. Assessment in Art and Design subjects is difficult: the skills involved in arts subjects are themselves complex, and furthermore "there exist many different conceptions of these skills" (Haanstra et al., 2015, p.413). Haanstra et al. go as far as to claim there is "no consensus on educational standards in the arts" (Haanstra et al., 2015, p.413). The particular demands of assessment in arts generally mean that the "forms and models of assessment particular to other areas of learning" do not transfer satisfactorily to Art and Design subjects

1. The vast majority of examination centres are schools or colleges.
2. The Joint Council for Qualifications (JCQ) is a membership organisation comprising the largest qualification providers in the UK. One of its aims is to provide common administrative arrangements for examinations.

(Eça, 2002, p.1). A consequence of this is that processes and concepts to do with evaluating assessment quality in other areas of learning also do not transfer directly to Art and Design subjects.

The Art and Design qualification used in this study contained two tasks, and could comprise of a variety of different art forms (e.g., annotated sketchbooks, mounted sheets, maquettes, prototypes, scale models, or written work). The candidates' non-exam assessment work (their submission/submitted work) were marked by the candidates' teachers. A sample, specified by the awarding organisation, of each centres' candidates' submissions, was then submitted for external moderation.

Previous moderation studies

Crisp (2017) used a *think aloud* method and moderator interviews to investigate moderation of General Certificate of Secondary Education (GCSE) assessments in English/English Literature, Geography, and Information and Communication Technology. Crisp described two groups of cognitive processes involved in moderation: (a) "reading and comprehending" the candidate work, and (b) making "evaluative judgments of quality" (Crisp, 2017, p.34). In terms of reading and comprehending, Crisp found that, in comparison with teachers, moderators were "more likely to make strategic choices about the level of detail in which they read different parts of students' submitted work" (p.34). Nevertheless, Crisp concluded that "the process of reading and understanding … appears to be unproblematic". This is in contrast to the evaluative processes, which are "rather more complex" (p.34).

The subprocesses that Crisp (2017) identified within the evaluative processes included attending to and evaluating features of candidates' work in relation to the marking criteria, indicating an analytic approach (p.34). However, Crisp also found that "most moderators appeared to apply 'configurational' processes in parallel", whereby "overall judgments are made directly and then checked against specific criteria" (pp.34–35). Reassuringly, Crisp found "no evidence of attention being paid to inappropriate features" of candidates' work and concluded that there was "no evidence of bias in judgments" (p.34). There was some evidence of affective reactions to candidates' work, but "these did not seem to influence judgments" (p.34).There was also evidence of comparative processes, these included comparison of a candidate's work to work by other candidates, comparison of candidates' work to other examples from the same candidate, and a tendency to arrange candidates' work in mark order.

The study by Cuff (2017) also used think aloud and interview methods. He used four specifications: GCSE History, GCSE English, GCSE Business Studies, and a Level 3 Extended Project Qualification. The aim of the research was to focus in greater depth on the cognitive processes involved in moderation, and on how moderators used possible supporting resources. Cuff noted that, in terms of the overall series of steps identified, "Encouragingly, these … align well with those reported by Crisp" (p.8). Many of the subprocesses that Cuff (2017) identified also aligned well with the details reported by Crisp (2017). In terms of resources, Cuff found that when reading the work, moderators formed impressions based on the marking criteria, previously moderated candidates, their understanding of the grade levels, and teachers' annotations written on the work. Crisp (2017) also found that moderators made use of annotations and the marking criteria when evaluating candidates' work.

Cuff found that "several aspects of the current findings suggest risks of confirmation biases in moderators' judgments" (p.35), which appears to contrast with the conclusions of Crisp (2017). The aspects that Cuff identified as potential sources of bias were an "anchor-and-adjustment" approach to adjusting marks (i.e., assuming the marks given by the centre were correct, unless shown otherwise), and the influence of moderators' initial impressions on their later judgements, even if the moderators themselves did not believe their judgements to have been affected.

Cuff (2017) recommended further research to confirm whether his findings applied "across a range of different contexts or where differences may exist (and why)" (p.37). To this end, we sought to explore the moderation process in terms of moderators' cognitive process and resources drawn on when making judgements about Art and Design submissions. Findings can contribute to the overarching moderation picture and help inform future training and moderation practice.

Method

The artwork used in the study was candidates' work submitted for an Art and Design qualification for 14–19 year olds. The Art and Design qualification contained two tasks: a Portfolio (worth 60%), and an externally specified Set Task (worth 40%). Both tasks were internally assessed by the centre and externally moderated. There were four Assessment Objectives (AOs) which were weighted evenly within each task. Five areas of study were available to candidates: Fine Art, Graphic Communication, Photography, Textile Design, and Three-Dimensional Design. Submitted work had to be in an appropriate format for the area of study and could take the form of, for example, annotated sketchbooks, mounted sheets, maquettes, prototypes, scale models or written work. The assessment of artwork for this qualification required holistic consideration of each candidate's submission, with marks assigned to each task.

The submitted work was sent to a central location where both the "live" moderation and this study took place. The researchers attended the standardisation meeting and observed some live moderation to enable us to mirror the live conditions as much possible. This study was conducted under experimental conditions several weeks after live moderation; this was because we did not want to disrupt live processes, nor risk affecting candidates' outcomes.

The study participants (N=3) were recruited from the small pool of moderators who had moderated the qualification in 2017. The participants all had significant teaching experience (15+ years) in Art and Design, and had taught Entry Level, GCSE, and General Certificate of Education Advanced Level (GCE A Level). They had all been or were currently Heads of Department and, at the time the study was conducted, all the participants held senior moderating positions.

Work from four centres was chosen, with each sample containing work from between two and eight candidates. The participants were instructed to moderate centre work in the same way that they would have done in the Summer 2017 session, using the evidence of candidates' work and resources that were available to them. Moderation was conducted for one centre at a time. The moderation task was to determine whether the specified marking criteria had been satisfactorily applied. In essence, this meant assessing whether the rank order of the centre sample was correct, and whether the marks given to the candidates' submissions were acceptable or would require adjustment. Participants were asked to record their marks and notes, as they would normally, and then write a report for the centre.

During moderation, a concurrent think aloud method was used.

The aim was to provide insights into the cognitive processes underpinning a specific activity through a verbalisation procedure (van Someren, Barnard, & Sandberg, 1994). The main advantage of this approach was that it provided researchers with additional information that would not be available through observation alone.

Prior to moderating, the participants were given a familiarisation task to give them the opportunity to get used to the think aloud method. The participants were provided with documents which replicated the materials that they would have had access to in live moderation. They comprised: a booklet of photographs that had previously been taken by the research team to represent the displayed artwork the moderation team had used during standardisation, a standards booklet (reference guide for moderators containing candidates' submissions benchmarked from across the mark range), a mark sheet, a copy of the marking criteria, a copy of the recording sheet on which moderators make notes of their observations, and a copy of the centre report template. The mark sheet contained the original marks for each candidate grouped by centre (total mark and mark by AO for each task).

The participants were asked to conduct moderating activities for approximately 90 minutes. They were allowed to take breaks at any time and it was made clear that they did not have to complete moderation for all centres. Participant activity was recorded via Morae software (TechSmith, 2011) and was observed by two members of the research team. The researchers sat beyond the participant's immediate line of sight. They recorded any relevant activities using observation schedules and noted anything that would comprise part of the interview to take place later in the day.

To account for the possibility that some of the participants might be more effective at verbalising their thinking than others, a retrospective interview was conducted with each participant after moderation had been completed. This was audio recorded. The aims were to illuminate and expand on think aloud outcomes, to add some information about the participants' thought processes, and to validate the researchers' early interpretations of the data collected.

Analysis

The recordings of the participants' spoken thoughts and activities from the moderation sessions were loaded into MAXQDA (VERBI Software, 2017). The research team familiarised themselves with the recordings by watching them and aligning them with the observation schedules coded during the observation period. An initial coding framework was developed with the aim of capturing the key activities; four categories of participant activity were identified:

1. Judgements about a candidate's level/mark;
2. Reference to documents;
3. Movements from submission to submission; and
4. Movements within a submission.

These broad categories were subdivided into several subthemes. For example, in the *Movements within a submission* category, the subthemes included leafing through work, leafing through work then focusing on one image, observation[3], speeding/leafing through a sketch book, consideration[4], lift up/bend down and touch/rotate work. The coding scheme was tested and refined. The researchers then double-coded (non-blind) all the data produced by the participants. This aimed to ensure consistency of application of the coding framework. Any disagreements between researchers were discussed and addressed. Typically, this took the form of a missed code. Within MAXQDA, it was possible to designate how long each coded activity lasted.

The interviews were first transcribed in MAXQDA, then analysed thematically.

Development of the process model

From the coded recordings of the moderation sessions, we developed a process model to describe how the participants carried out moderation. Firstly, for each separate moderation session, the codes described above were mapped against time. The timelines covered the period from the start of the moderation session (no work had yet been viewed and no other preparation work had yet begun), to the point at which the participant was ready to write the moderation report. Simplified timelines were also created (from the fully coded timelines) to show the candidate work each participant was viewing throughout moderation. For illustration, an annotated simplified timeline is shown in Figure 1.

3. Observation refers to comments on a candidate's/multiple candidates' use of styles, techniques and artists – all made without judgement
4. Consideration refers to comments with some element of judgement about the quality or realisation of the work of a candidate/multiple candidates. It denotes deeper engagement with the submission.

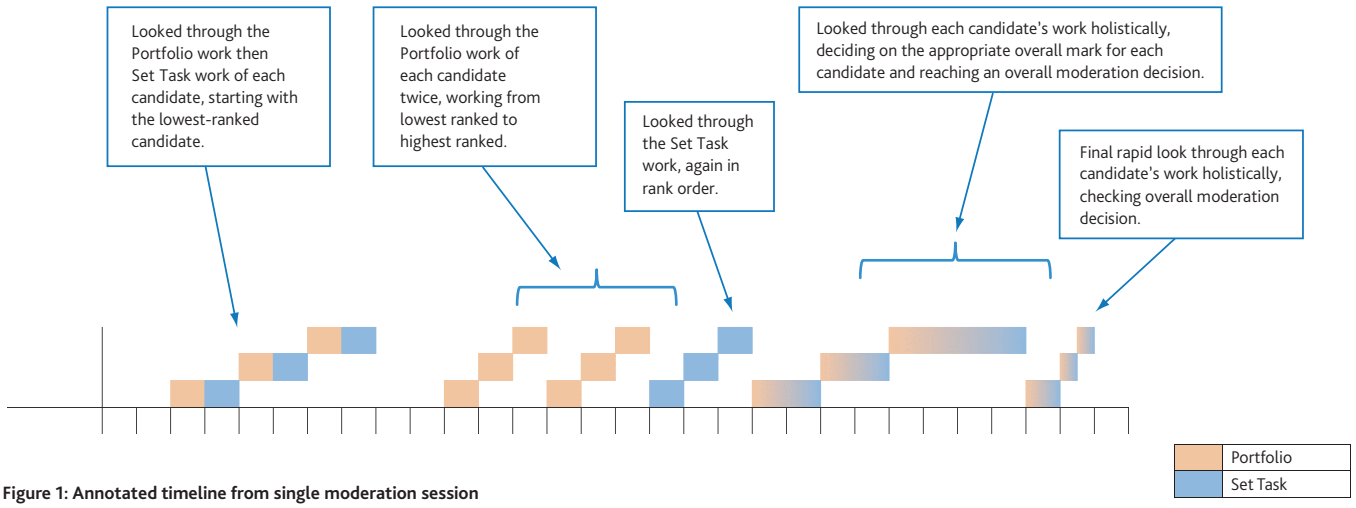

Figure 1: Annotated timeline from single moderation session

Secondly, the coded timelines were then compared to the moderation models of Crisp (2017) and Cuff (2017), to see how far these existing models were applicable. We found that while these existing models did not fully or accurately describe the observed moderation, certain features were evident. Although we could not use either model "as is", the models proved to be a useful frame on which to develop our artwork moderation model.

Thirdly, the coded timelines for each moderation session were synthesised into annotated diagrams, which were further combined into a single overall representation of moderation. The different stages of the process model were developed by: identifying repeated and systematic occurrences of distinctive combinations of cognitive processes, physical activities, and resources. For example, the familiarisation stage was characterised by rapidly looking through candidates' artwork throughout the sample, with high-level observations about themes, artists and techniques; some inferences about the course and/or centre; initial evaluation of the sample; and using the mark sheet to arrange work in mark order.

While there were some variations in the order in which the participants viewed work, the cognitive processes, physical activities, and use of resources formed coherent and identifiable stages that were common across all three participants – the differences in order did not necessitate separate process models for each participant.

Findings

Figure 2 shows the process model; we start by describing the model and then explore the differences in activities and cognitive processes that lie behind the moderation stages. The model is arranged in four columns: The first column shows the overall stages of moderation: orientation and preparation, familiarisation, investigation, reaching a moderation decision, and report writing. The second column shows the observed activities associated with each stage. The third column shows the cognitive processes associated with the stage (inferred from think aloud data), and the final column shows the resources drawn upon at each point.

Throughout the model, dotted lines indicate elements that varied among the three participants. For example, for two participants, the activity of setting up the moderation recording sheet occurred during the orientation and preparation stage, but one participant set up the recording sheet only after the familiarisation stage.

The next sections describe the stages of the model; inserted quotations illustrate activities that were typical in the different stages.

Orientation and preparation

The first stage of the observed moderation was an orientation and preparation stage. During this stage, the participants orientated themselves to high-level features of the centre and sample. In particular, from looking at the mark sheet, they noted how many candidates were in the sample, the marks given by the centre (centre marks), the rank order of candidates, and any unusual features. They also determined an order for the physical layout of candidates' work.

> *The candidates in this line up–and its centre XXXXX–there are three candidates. And … [writing down the candidate numbers] the marks are–the total marks are 64, 51, and 40.* (Participant C)

Two of the three participants prepared the recording sheet during the

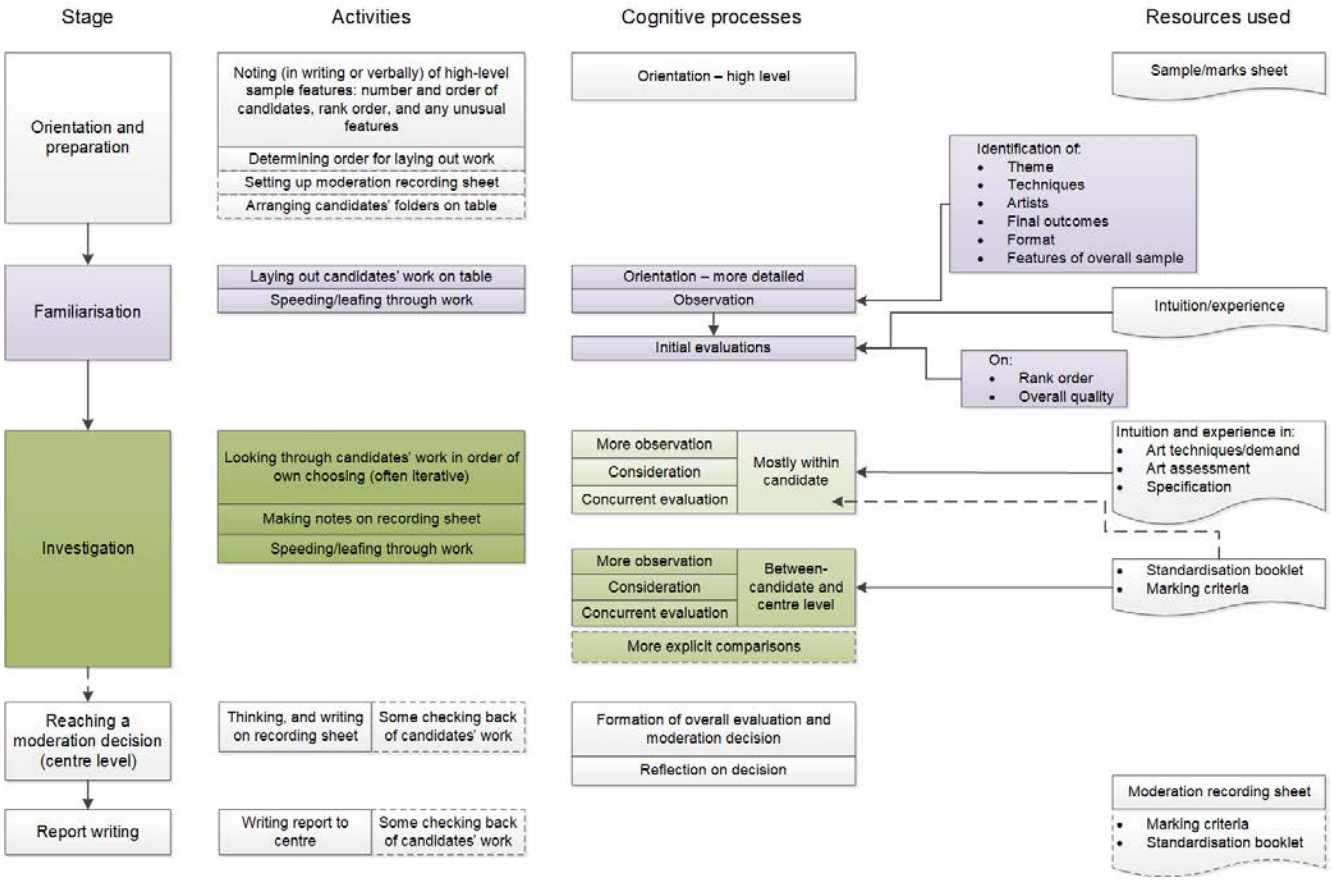

Figure 2: Artwork moderation process model

orientation and preparation stage. They added the centre details to the blank recording sheet, and transferred the marks from the mark sheet.

Familiarisation

The key activity carried out in the familiarisation stage was laying out candidates' submitted artwork on the table, in the order determined in the orientation and preparation stage. The artwork was spread out so that individual sheets of work were all visible, wherever possible. The participants each used a slightly different layout, but all were based on the rank order of candidates' centre marks. Each participant's chosen layout maintained a distinction between the Portfolio work and Set Task work of each candidate.

Whilst laying out the candidates' work, the participants leafed through candidates' work fairly rapidly. During this activity, the participants made observations on the theme(s) of the submitted work, the artists that the candidate showed evidence of studying, and on the techniques demonstrated by the candidate. The participants identified the format of the submitted work, and the "final piece" (final outcome) within the Portfolios and Set Tasks.

Besides observations about particular candidates or particular pieces of work, the participants made observations about features of the centre's sample overall. These observations included comments on the similarities and differences amongst candidates:

> *So, they've all done the same project …* (Participant B)

> *And what's interesting already by candidate … number two, I can see that–in terms of the procedure–of how the centre has got their candidates to produce their work, that the Portfolio is a study sheet with a sketchbook, and … the externally Set Task is … the externally Set Task is just using study sheets.* (Participant C)

In some cases, the participants needed to revise initial expectations and judgements as they progressed through the sample. In this quotation, the participant had (from the first two candidates' folders) concluded that all candidates in the sample had completed Portfolio work on the same theme, but revised this upon first looking at the next candidate's folder:

> *And, I bet this is space again! … Now, let's have a look at this [laying out candidate's work on the table]. Oh no! Mechanisms. So this candidate has worked–probably slightly more independently, chosen their own … I'm just supposing.* (Participant A)

Another distinctive cognitive activity that characterised the familiarisation stage was orientation to the submitted work. In contrast to the orientation activity within the *Orientation and preparation* stage (which could be summed up as orientation to *whom* the participants were moderating), the orientation within the familiarisation stage equipped the participants to know *what* they were moderating. The participants familiarised themselves with what had been submitted, and, crucially, how the individual materials within candidates' folders related (or did not relate) to the course requirements and marking criteria. Having briefly viewed the whole sample's Portfolio work, one participant explained:

> *At this stage, what I do is just make some preliminary notes, based on what I can see. So, we've got two distinct projects… and the first one is of 3D model making, which is about letters …* (Participant B)

In the following quotation, the participant explains how the laying out of the candidates' work has resulted in a physical overview, enabling location of work relating to specific AOs, at the same time as seeing this work in relation to the "final outcome" work:

> *I can see clearly now where the individual AOs are, and how they relate to each other. So I can see the 'Explore', just in this overview, and how it's impacting on the outcomes.* (Participant C)

The final cognitive activity that took place within the familiarisation stage was initial evaluation. Drawing on their intuition and experience (not yet on the formal resources of marking criteria or standardisation booklets), the participants made observations on the quality of submitted work. The participants themselves used the phrase 'first impressions', emphasising that these judgements were open to revision at a later point, and differed from the final professional judgements that the moderation process worked towards:

> *First impressions are that it looks strong.* (Participant B)

> *First impressions are that it looks terr–it looks under-marked …* (Participant A)

> *It's got a mark of 51–Level 1–and this is definitely higher than Level 1.* (Participant A)

> *Some good lettering there … This is definitely more than just into Level 2.* (Participant A)

During the familiarisation stage, the participants carried out an initial review of the rank order through laying out and looking at candidates' work in order of mark. None of the participants commented during the familiarisation stage that the centre's rank order was incorrect. However, if the participants' initial impressions of the submitted work were incompatible with the centre's rank order, it would necessarily have become apparent by the end of the familiarisation stage.

One participant described the activities of the familiarisation stage as indicating the "flavour" of what had been submitted for moderation:

> *By walking up and down the line I am actually registering the work, the standards of the work, I am actually getting that overview immediately and it means then, if you like, that when I come back to go through in terms of centre marks and the assessment, I've already got the flavour of what I'm looking at.* (Participant C)

This quotation conveys the participant's sense that the familiarisation stage provided the foundations for the later stages of moderation, during which detailed consideration of "centre marks and the assessment" would occur. A quotation from a different participant, at the end of laying out the whole sample's submitted work, similarly underlines the role of the familiarisation stage as a preparatory stage, and separate from the "actual" moderation process:

> *So, that's that. So, moderating. We'll get into moderating now.* (Participant A)

This distinction on the part of the participants supports the separation of the familiarisation and investigation stages in the process model. Although there was overlap in the types of activity observed within the familiarisation and investigation stages, the details and purposes of the activities observed were different, and the participants themselves appeared to consider them as separate stages.

Investigation

The primary activity of the investigation stage was looking through candidates' submitted work. The participants continued to make observations about the work, specifically, more extensive and more detailed observations on the same set of characteristics noted in the familiarisation stage. The participants also considered the quality of work, made evaluations of the work in comparison to the centre marks, and made explicit comparisons between examples of candidates' work, including those in the standardisation booklet. The participants made notes on the moderation recording sheet.

> *And this is a far more substantial portfolio of Set Work [Set Task]. He's looked at [Artist name 1], he's done mechanisms, and also … [Artist name 2].* (Participant A)

> *I would say that in terms of the sophistication, the confident experimentation, I would say … the quality of some of this drawing is very strong.* (Participant B)

> *Well actually, this set of work … doesn't quite have the achievement of the Portfolio. … And looking at our Set Work standards, and I'm firstly comparing–excuse me–to the 28 … It's slightly better than the 28, doesn't get to the 33 in our standards booklet.* (Participant C)

The first clear resource drawn upon during the investigation stage was participants' intuition and experience. Their knowledge and experience of art techniques (particularly the skill required to achieve particular outcomes), art assessment, and the course requirements were drawn upon frequently. The external resources drawn upon during the investigation stage were the marking criteria, and the standardisation booklet containing examples of candidates' work at particular levels, as referred to by Participant C above.

> *And that could go up. I'm going to–I'm just going to go and check my Level 2 criteria …* (Participant A)

> *Looking at the exemplar on 52, I would say this is stronger than the 52.* (Participant B)

The investigation stage was the most complex and lengthy of the moderation processes observed. The participants varied in terms of the order in which they looked through candidates' work, the precise point at which they drew on external resources, and the number of times they viewed the total sample. Participants viewed the work in rank order – one from the highest ranked candidate and two from the lowest. For two participants, the investigation stage was a highly iterative stage, and they each looked through the whole sample multiple times.

For all the participants, there was a gradual shift in the content of the considerations and evaluations of candidates' work that occurred during the investigation stage. As Figure 2 suggests, cognitive activity in the earlier stages of the investigation stage tended to focus on particular candidates. Towards the end of the investigation stage, cognitive activity more frequently focused on between-candidate comparison and centre-level consideration. Overall, a characteristic of the investigation stage was that the participants moved from detailed consideration of particular candidates' work, towards a point where they were ready to reach a moderation decision on the centre overall.

Reaching a moderation decision

The stage of reaching a moderation decision was the culmination of the investigation stage. The cognitive activity of this stage was forming an overall evaluation of the centre's marking, and reaching a decision on which marks to recommend.

The participants differed in the length of time they spent at this stage. For Participant C, the overall evaluation had been built up during the investigation stage – to the extent that reaching a moderation decision consisted of little more than stating and writing down the overall judgment. For Participants A and B, more time was taken, and there was a more sustained period of checking or reflecting on the decision. Both Participant A and Participant B looked through the entire sample again (Participant A did so twice) during the process of reaching and confirming their overall moderation decision.

> *The effect of the moderation so far is that we're moving some of the marks from where they were in Level 1 up to Level 2.* (Participant B)

Report writing

In the final moderation stage, the participants wrote a moderation report for the centre. Observation revealed that all the participants drew upon their moderation recording sheet to write this report, and some of the participants also referred to the marking criteria and standardisation booklet.

Aside from the activity of writing, the other activity observed during the report writing stage was some checking back to specific aspects of the submitted work. These checks were typically brief (sometimes just a glance) and often served to confirm a specific aspect of submitted work that the participant had referred to on the moderation recording sheet.

Discussion

Before we discuss the findings, it should be noted that there are some limitations to this research. First, it is necessary to exercise caution in generalising the findings given the small sample size. Although only three participants were involved, this was 60 per cent of the (small) population that moderated the target qualification. With only three participants, it is possible that individual differences could account for some of the findings. The study attempted to replicate live moderation as much as possible. Participants, however, were fully aware that this was a research exercise and their decisions would not contribute to candidates' results. In addition, the think aloud method used in the present research might potentially have increased the cognitive load for the participants, which might have influenced their moderation activities. However, it should be noted that when the participants were asked if the think aloud method disrupted their moderation activities, they perceived that it did not do so. This is in line with research by Greatorex and Suto (2008), in their study of the marking processes of 12 GCSE examiners. They found no relation between the type of items being marked (as a proxy for task difficulty) and the perceived ease with which the participants were able to think aloud.

To summarise our results, we found that the observed moderation process began with an orientation and preparation stage, followed by a familiarisation stage and then featured a lengthy investigation stage. This was followed by a stage in which a moderation decision was reached, and then finally a report writing stage. In terms of cognitive processes, the participants oriented themselves to the moderation task, made observations, considered and evaluated candidates' work, made explicit comparisons, formed overall evaluations and a moderation decision, and reflected on this decision. All of the participants made

observations and considerations about multiple candidates within the first ten minutes of moderation, and throughout the investigation stage.

As stated earlier, we compared our data and timelines to the models of Crisp (2017) and Cuff (2017) and we found these models did not fully or accurately describe the observed moderation. Certain aspects of the models, however, were evident in the observed moderation of artwork. This was either directly, or when a different but parallel process was substituted; for example, the processes of scanning and reading in the Crisp (2017) and Cuff (2017) models could be replaced by qualitatively different forms of looking at candidates' artwork (e.g., leafing and consideration).

We found the sequence of activity in the observed moderation sessions most closely resembled the structure of the Crisp (2017) model. In contrast, the specific cognitive processes identified in the Cuff (2017) model more precisely described the subprocesses we observed in the participants' moderation activities. Not every process identified by Cuff (2017) was included in the art moderation model, and we found observed processes in the think aloud data that Cuff (2017) had not included that needed to be added. For example, the participants in this study started comparing submissions from the start of the moderation session.

Similarly to the Cuff (2017) model, we incorporated resources into the art moderation model. We found that all participants used the additional resources, particularly making reference to the standardisation booklet and marking criteria. This is in contrast to Cuff (2017), who found that some moderators relied solely on internalised standards.

We found some differences to both the Crisp (2017) and Cuff (2017) models. Firstly, the participants in this study moved through different candidates' work repeatedly in a cyclical fashion, building up an impression of submission quality. We hypothesise that this was because the work was on full view and did not require detailed or lengthy reading in order to get an impression of quality. Secondly the participants in this study made little, if any, mention of the teacher annotations, whereas this was prominent in the previous research. This could be due to the moderators not needing to read long passages of text and so not needing the hints provided by the annotations.

As the complex structure of both our art process model and the Cuff (2017), model highlights, the moderation process combines many cognitive processes and draws on many resources. That we needed to develop a new model, rather than use one of the existing ones, could be due to the nature of the submissions. The Crisp (2017) and Cuff (2017) models were developed for the moderation of written work, whereas our study used artwork which contained very little text. The overall moderation process should be the same for any subject overseen by the JCQ as it is subject to agreed and documented procedures. Indeed, we found this to be the case for Art and Design, and the overall sequence followed that described by Crisp (2017). We found, however, that the subelements and their interaction with resources did differ for art, suggesting that subject-specific differences may exist at the sublevels.

What all three models emphasise, however, is the iterative and evaluative nature of the moderation process and its focus on quality control. Similarly to Cuff (2017), we found that aspects of the findings support the validity of the moderation process. We too found that the participants followed similar stages, made reference to the marking criteria, focused on appropriate features in the submissions, were mindful of being fair to candidates, and displayed thoroughness in making their judgements.

References

Crisp, V. (2017). The judgement processes involved in the moderation of teacher-assessed projects. *Oxford Review of Education, 43*(1), 19–37.

Cuff, B. (2017). *An exploratory investigation into how moderators of non-examined assessments make their judgements*. (Ofqual/17/6252). Coventry: Ofqual.

Eça, T. (2002). *A conceptual framework for art and design external assessment*. Paper presented at the European Conference on Educational Research, University of Lisbon, Portugal.

Gill, T. (2015). The moderation of coursework and controlled assessment: A summary. *Research Matters: A Cambridge Assessment publication, 19*, 2–6.

Greatorex, J., & Suto, W. M. I. (2008). What do GCSE examiners think of "thinking aloud"? Findings from an exploratory study. *Educational Research, 50*(4), 319–331.

Gruber, D. D., & Hobbs, J. A. (2002). Historical Analysis of Assessment in Art Education. *Art Education, 55*(6), 12–17.

Haanstra, F., Damen, M.-L., Groenendijk, T., & van Boxtel, C. (2015). A review of assessment instruments in arts education. In S. Schonmann (Ed.), *International Yearbook for Research in Arts Education. The wisdom of the many: Key issues in arts education* (pp. 413–418). New York: Waxman.

Herpin, S. A., Washington, A. Q., & Li, J. (2011). *Improving the Assessment of Student Learning in the Arts – State of the Field and Recommendations*. Washington: The National Endowment for the Arts.

Joint Council for Qualifications. (2018). *Instructions for conducting non-examination assessments (new GCE & GCSE specifications)*. Retrieved from https://www.jcq.org.uk/exams-office/non-examination-assessments/instructions-for-conducting-non-examination-assessments

Mason, R., Steers, J., Bedford, D., & McCabe, C. (2005). The effect of formal assessment on secondary school Art and Design education: a systematic description of empirical studies. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

TechSmith. (2011). Morae Recorder [Computer software]. Retrieved from https://www.techsmith.com/morae-features.html.

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modelling cognitive processes*. London, UK: Academic Press.

VERBI Software. (2017). MAXQDA: Software for qualitative data analysis [Computer software]. Retrieved from https://www.maxqda.com/.