An exploratory investigation into how moderators of non-examined assessments make their judgements

July 2017

Ofqual/17/6252

Author

This report was written by Benjamin M. P. Cuff, from Ofqual's Strategy, Risk and Research directorate.

Contents

1 Executive summary

Non-examined assessments (NEAs) are used to assess student competencies not easily accessed via written exams. As NEAs are usually marked internally by centres, exam boards are required by Ofqual to moderate the marks awarded by centres, to ensure that the mark scheme has been appropriately and consistently applied. Although the general process of moderation is relatively well documented and understood (eg sampling, mark adjustments, etc.), the process through which moderators make their decisions is less well known (ie the thought processes involved). The purpose of this research was to investigate moderators' decision making processes in order to identify what constitutes current practice in this area, in order to help us determine what might constitute best practice.

A qualitative study using a mixture of think-aloud and traditional interviewing techniques was conducted using a sample of 10 moderators from four exam boards and subject areas. A particular focus of these discussions was on how moderators make use of the various mental and physical resources available to them. Through analysing this interview data, a model of the moderation process was developed. This model can be summarised as follows:

- Before reading each script, moderators begin by forming various expectations about the likely quality of the work produced. These expectations are largely based upon information given in any documentation sent to them by the centre. For example, certain expectations are made based upon knowledge of the centre (eg its reputation) and the candidate. Expectations are also based upon the marks that had been awarded by the centre, and upon the rank order that the centre had placed scripts into (ie lower marked scripts are expected to be of a lower quality than higher marked scripts). Some expectations are also based upon the various comments that the marking teacher had made about the work.

- When looking at the main report, but before reading it through, moderators form various first impressions. These first impressions are based upon surface features of the work (eg the title/the length of various sections) and upon spelling mistakes or on whether any rules of the assignment brief have been broken.

- When reading the main body of the text, moderators develop their impressions of script quality by focussing on the quality of the writing, and how the work compares to descriptions given in the mark scheme. Some moderators find it helpful to compare the script under review to those that had been previously moderated, and to their understanding of different grade levels. Both of these help to frame moderators' thinking during this reading phase. Some moderators also find centres' annotations on the work helpful in terms of highlighting certain

aspects of the report that they may have otherwise missed and in terms of understanding why the centre had awarded a certain mark.

- Moderators make use of various ‘benchmark’ resources to help them make and evaluate final decisions for each script. For example, each script under review is often compared against the marking criteria, grade descriptors, and other scripts, to help make their decisions. Intuition or experience is another useful resource for some. Other considerations are taken into account when making and evaluating decisions, such as tolerance thresholds (ie the level of disagreement between the moderators’ and centres’ marks), the rank order of centre’s marks, centres’ comments, and fairness for the individual student and for the cohort overall.

The findings of this research foster a greater understanding of this important validation process in the assessment of learning outcomes. However, although several positive aspects of this process were apparent, certain potential risks to the validity of moderators’ judgments were identified. For example, elements of practice may lead to an increased likelihood of agreement with centres’ original marks (confirmatory bias). Other biases might also arise from the fact that centre/candidate information is not anonymised to moderators. The use of grade boundaries as a basis for making and evaluating decisions may also be problematic as these boundaries are subject to change during awarding. Further work is therefore needed to assess the degree of impact that these elements have on the validity of moderators’ judgments.

2 Introduction

Non-examined assessments (NEAs), such as coursework and controlled assessments, allow for the assessment of student competencies that are less easily accessible via externally assessed exams. These might include practical knowledge and skills (eg in the sciences), performances (eg in the performing arts), or any other knowledge, understanding, or skills that cannot be easily assessed in a written exam (eg speaking skills in modern foreign languages) (Ofqual, 2013). NEAs are usually marked internally by centres. Given the importance of maintaining common standards across different centres and assessment series, exam boards are required by Ofqual to externally moderate the marks awarded by centres.

The general process of moderation is fairly well documented, such as the sampling of scripts for moderation, and how post-moderation mark adjustments are calculated (eg see JCQ, 2016; Johnson, 2011; Ofqual, 2011, sec. 5 - For AO-specific documentation, see AQA, 2013; OCR, 2015; Pearson, n.d.; WJEC, 2015). However, we are only aware of one piece of existing research that explores how moderators actually make their decisions: Crisp (2017) produced a model of the moderation process, focussing upon the key stages of moderators' decision making (which were: orientation to the sample and determining the order for consideration; orientation to topic/title; initial scan; reading and concurrent evaluations; overall evaluation, mark consideration and mark decision; reflection on mark; reviewing mark differences and making a decision about whether to accept the school's marks). In the current research, I place greater focus onto the thought processes involved in each stage, and more specifically, how moderators make use of the various mental and physical resources available to them. The purpose of this was to foster a better understanding of what constitutes current practice in this area, which may help us determine what constitutes best practice, and what improvements to the system might be suggested.

In this report, I shall begin with a brief description of current UK practice, followed by the presentation of a qualitative study undertaken to improve our understanding of this important validation process in the assessment of learning outcomes.

2.1 Moderation of NEAs in the UK

In current practice, a form of 'moderation by inspection' is used in the UK for general qualifications (ie GCSEs and A levels). In essence, this means that external moderators (employed by the exam boards) evaluate the marks awarded by centres, to assess the consistency and appropriateness of the application of the marking criteria (Daly et al., 2011). This is also known as a 'social moderation' approach;

purely statistical moderation approaches[1] are not used in the UK, and so interested readers are directed elsewhere (eg Williamson, 2016; Wilmut & Tuson, 2005). For reviews of the history of non-examined assessments in the UK, see Johnson (2011) and Ofqual (2013). Ofqual (2015) describes some subject-specific regulations.

Many NEAs in the UK are marked according to level descriptors (rather than points-based mark schemes), meaning that moderators must use a degree of professional judgment when making decisions. Before 'live' moderation begins (ie the actual moderation of work submitted for consideration towards a qualification), moderators are required to attend a standardisation meeting, led by the principal moderator[2], to foster a shared understanding of the marking criteria. This meeting typically involves individual and group scrutiny of a number of non-live scripts that have been pre-selected by the principal moderator (known as 'standardisation scripts'). Sometime after this meeting, live materials are posted to moderators directly from the centre. Scripts are not handled or anonymised first by the exam boards.

For each centre, moderators begin by reviewing just a sample of the work that was sent to them. If they agree with the centre's marks at this stage (within a specified tolerance), then the centre's marks are approved, and no further moderation action is taken. If the moderator disagrees with any[3] of the centre's marks outside of a certain tolerance, they are required to review a further sample of work. Where consistent differences between the moderator's and centre's marks exist, an adjustment is applied to all candidates' marks from that centre, and not just those in the moderation sample. The purpose of moderation, therefore, is not to remark individual scripts, but rather to align standards across different centres. Mark adjustments are calculated or determined by the exam boards, rather than the moderators themselves, and most boards make this calculation on the basis of a regression line of the relationship between the moderator's and centre's marks (see Pearson, n.d.,

[1] Eg scores on another assessment may be used to 'calibrate' standards on the assessment being moderated.

[2] Moderation teams are usually led by a 'principal moderator' and 'assistant principle moderators'. 'Team leaders' are responsible for smaller teams of 'assistant moderators'. For simplicity, the term 'moderator' will be used throughout this report, but specific roles shall be made apparent where pertinent to do so.

[3] Some exam boards allow moderators to declare one script per sample to be an 'outlier', which is then ignored for decision making purposes, meaning that two scripts out of tolerance are needed to trigger the next stage in the process.

for a simple explanation of how this works). In effect, this method aims to preserve the centre's rank order of candidates, but adjusts all marks to fall in line with national standards. In cases where the pattern of a centre's marks is substantially or inconsistently different to the moderator's (ie where a centre has marked unreliably), all scripts from that centre may need to be moderated or re-marked.

To summarise, there are 3 possible outcomes of this process:

1. the moderator agrees with the marks awarded by the centre (within a specified tolerance), and so the centre's marks are accepted;
2. the moderator disagrees with the centre's marks (out of tolerance), and so marks are adjusted via the aforementioned calculations; and
3. the moderator believes that marking has been carried out in a particularly inconsistent manner, and all scripts may need to be remarked.

3 Study aims and methodology

Full details of the method can be found in the appendix. In the interests of brevity, only a summary of key points shall be presented here.

As previously discussed, although we have a good understanding of what decisions moderators can make, and what actions may be result from these decisions, we have a relatively limited understanding of how moderators actually make their judgements. A qualitative study was therefore carried out to further explore moderators' decision making, focussing upon what physical and/or mental resources they draw upon to help them in their work. Such explorations can grant a better understanding of current practice, so that we might better understand what constitutes best practice, and what improvements to the system might be made.

A mixture of retrospective 'think-aloud' and more traditional interviewing methods were employed to gain insights from 10 moderators of 4 different specifications from 4 different exam boards. These were GCSE history, GCSE English, GCSE business studies, and a Level 3 (equivalent to A level) extended project qualification. The purpose of this design was not to identify any differences between these subjects/exam boards/levels of study, but rather to gather insights from a range of different moderators. Where differences were identified, however, these shall be made clear in the relevant discussions within the results section.

Moderators were interviewed in two waves, reflecting the two main moderation windows: 5 were seen in June, and 5 were seen in November. Once each wave had been completed, audio recordings were transcribed by an external transcription company, and these transcripts were coded and analysed using thematic analysis. Findings from Wave 1 were used to develop the interview schedule for Wave 2, to verify and further explore any hypotheses made during the first wave (otherwise known as a ‘grounded theory’ approach – eg Strauss & Corbin, 1994). The identification of any differences between June and November sessions was not one of the aims of this research.

4 Results

After reviewing, and re-reviewing the transcripts and analytical codes, it became apparent that moderators were following a similar overall series of steps in their work. The model of the decision making process underwent several iterations throughout the analysis of findings, but the final model is shown in Figure 1. This outlines the common series of steps that moderators took for each sample, with the middle 5 steps being repeated for each script within each sample. Some differences were identified between moderators in the extent to which they used some of the mental and/or physical resources available to them. For the remainder of the results section, high-level headings (5.1, 5.2, etc.) will represent the overall series of steps taken, and each subheading (5.1.1, 5.1.2, etc.) shall represent a mental or physical resource that was used to help moderators progress through each step.

In brief, moderators begin by preparing the sample and relevant materials. They then reviewed centre documentation and scanned surface features of the work to form expectations and first impressions. Once they began to review the work itself, they read and evaluated the main body of the script, making and evaluating overall decisions at the end. Once all scripts within the sample had been reviewed, final decisions for the centre were made. Encouragingly, these series of steps appear to align well with those reported by Crisp (2017).

Findings indicated that there were few systematic differences between moderators from different boards/subjects (at least not that could be perceived in this sample). Therefore, to avoid compromising anonymity (which may have been put at risk with such small sample sizes per group), moderators’ affiliations are anonymised in the following discussions. Individuals shall be mostly identified by their participant number. The exceptions to this are instances where differences between boards/subjects become pertinent to the discussion, or when the particular job role needs to be identified, at which point the labelling scheme will be changed to avoid participant numbers being linked with particular affiliations.

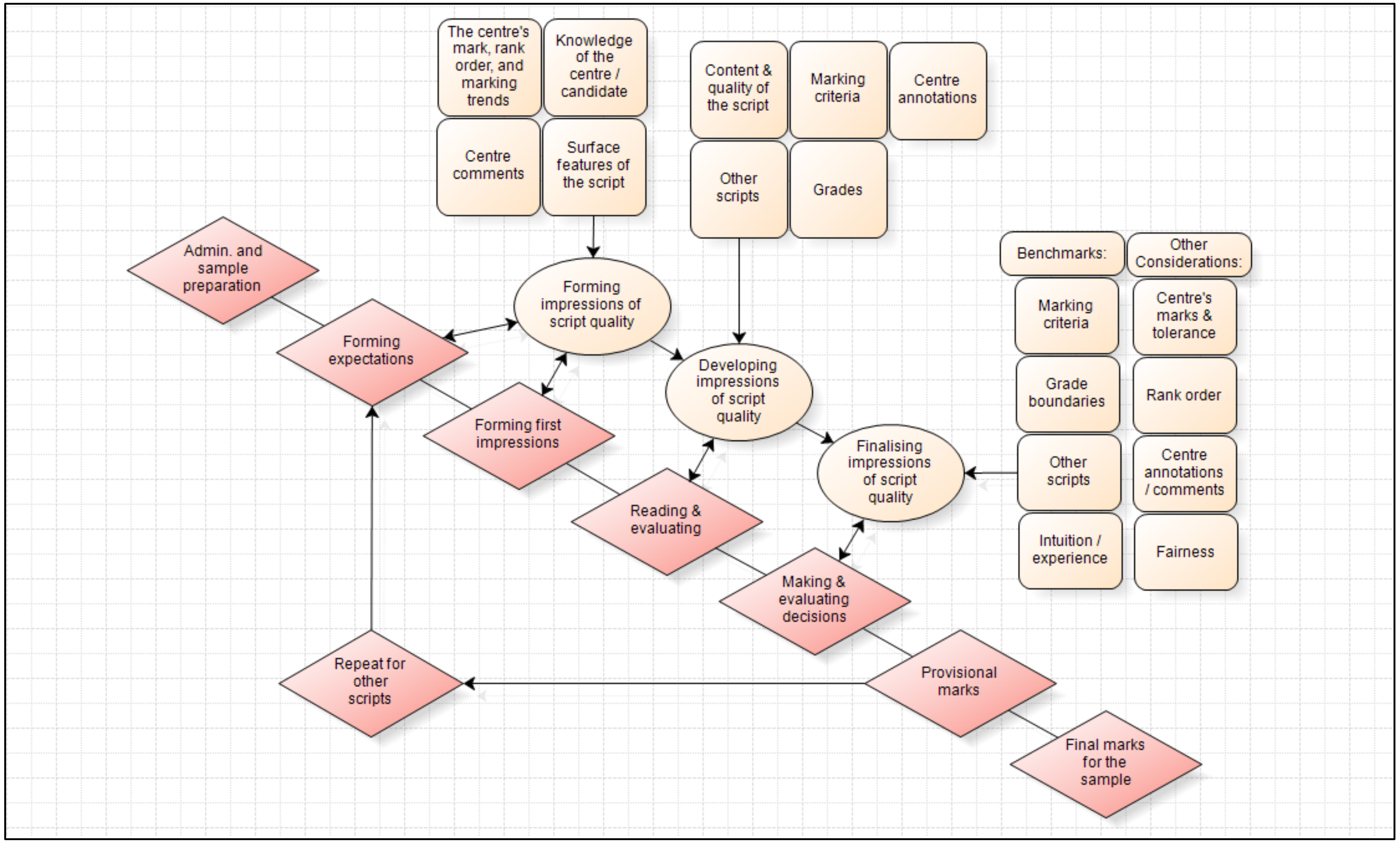
The centre's mark, rank order, and marking trends
Knowledge of the centre / candidate
Centre comments
Surface features of the script
Content & quality of the script
Marking criteria
Centre annotations
Other scripts
Grades
Admin. and sample preparation
Forming expectations
Forming impressions of script quality
Forming first impressions
Developing impressions of script quality
Reading & evaluating
Finalising impressions of script quality
Making & evaluating decisions
Benchmarks:
Other Considerations:
Marking criteria
Centre's marks & tolerance
Grade boundaries
Rank order
Other scripts
Centre annotations / comments
Intuition / experience
Fairness
Repeat for other scripts
Provisional marks
Final marks for the sample

Figure 1. Model of the decision making process for each sample.

Notes. Red diamonds represent steps taken (ie actions), yellow ovals represent the progression of moderators' impressions of script quality, and yellow squares represent mental or physical resources. All identified resources are shown, but the use of these resources differed according to context.

4.1 Admin and sample preparation

Before moderators read each script, there were a number of administrative checks that needed to be completed, such as checking whether the correct number of materials had been sent by the centre, and whether the necessary forms had been signed. A number of additional materials were also prepared. Some moderators had physical copies of the marking criteria with them, although others felt able to do their work without a physical copy to hand (they felt that they had sufficiently internalised it). Some also had a number of 'standardisation scripts' with them (ie the scripts that had been deliberated upon during the standardisation meeting). Again, some moderators did not have these physically to hand, as they felt that they had sufficiently internalised the benchmark standards.

As described in Section 2.1, moderators initially only look at a sample of the work that had been sent to them. Although the composition of this sample is usually determined electronically by the exam boards, one board allowed their moderators a degree of choice in script selection. These moderators were instructed to select the script with the highest mark (awarded by the centre), the script with the lowest mark, and any other three in between.

One moderator who was allowed such a choice found it helpful to select scripts close to what they expected the grade boundary marks to be (although grade boundaries for controlled assessments can change year-on-year, they often remain relatively stable). The same moderator explained how his/her exam board advised moderators not to choose any script that had been awarded a mark of below 20.

> *Moderator 3*
> It's nice to see what a C [grade script] looks like. So that's why I like to choose a C, because that's the difference between getting the grade that everyone's happy about and not getting it.
>
> [Interviewer: You said before that you ignore anything below 20 marks. What was the reason for that?]
>
> We're advised to… I've got a couple of pieces of work here. One's got two, which it's not worth looking at, is it, for two? One's got 10. How can you moderate a piece of work with 10?... And there's another one with 16 here, which I suppose could be looked at because 17 is an E, but once you do look at these pieces of work, invariably the marker is right.

In terms of deciding which order to work through the sample, moderators took one of two approaches. Three moderators simply reviewed the scripts in the order given to them (usually in alphabetical order, or by candidate number). The rest reordered the scripts according to the rank order of centre marks, and then worked through the sample in terms of this rank order (ie from highest mark to lowest mark). A slight variation of this method, taken by 2 moderators within this group, was to review the top script first, then the bottom script (to get a sense of the range of marks), and then work through the rank order from second highest to second lowest mark. Working through the rank order appeared to allow moderators to more easily identify any trends in marking, as one script served as a benchmark for the next.

> *Moderator 4*
> Some people maybe do them randomly, but I think it's quite difficult to see trends that way… So, personally, I always mark them in descending order.
>
> *Moderator 10*
> I usually start at the top… going down to the one who's got the mark below, then the mark below that. Because then they're falling into a pattern with each other. I couldn't then jump down to a 15 or something, because there'd be no relationship between that and the 45. But the 45 I can relate to the 40.

4.2 Forming expectations

After selecting a script to review, but before reading it, moderators were able to form expectations about the likely quality of the work, or of the marking by the centre, based upon the details given by the centre on their documentation. However, 2 moderators in the sample (each from different exam boards/subjects) noted that they tried not to form such expectations, and tried to remain as objective as possible prior to reading the work. It is perhaps worth noting that both of these were more senior moderators (a team leader and a principal moderator).

> *Team leader*
> As much as humanly possible I wouldn't prejudge… I suppose it's natural to some degree [to form some expectations], but based on experience it doesn't mean anything, so I wouldn't use that… to prejudge how it's going to be.
>
> *Principal Moderator*
> I try to go to it without any preconceptions at all… I try to, and I hope our moderators do, come to it without any kind of preconceptions there.

Nevertheless, most moderators did appear to form some expectations prior to reading the work. These were generally based upon their knowledge of the centre or information provided about the candidate (centre/candidate information is not anonymised for moderation), as well as the marks that the centre had awarded. Expectations were used by moderators to help frame their thinking when it came to reading the main body of the work, although of course any aspects of the script that broke those expectations could lead to a change in opinion.

4.2.1 Knowledge of the centre/candidate

Some moderators noted that the centre name, which would have been apparent on any documentation, may have had some influence on their expectations. Although several claimed that these expectations had no explicit bearing on their final decisions, the possibility for implicit biases should perhaps be noted (this shall be discussed further in the discussion section). One moderator also noted the possibility for bias caused by information provided about the student (ie the students’ names).

> *Moderator 4*
> Obviously, it’s none of my business, but it says ‘merit pupil referral unit’ [on the centre’s documentation]. So it might well suggest from that that it’s a non-specialist that’s been teaching them.
>
> *Moderator 10*
> There are obviously one or two schools that stand out, you know, because they're rather prominent national schools for one reason or another... It's very difficult not to have expectations about the school… You look down the mark sheet and you find that virtually every candidate is between 45 and 50 [out of 50]. Then you look at the name of the school and you can't help but think, oh yeah, OK.
>
> *Moderator 2*
> You know, maybe that’s something that could be changed; take the school names off… I’d have thought that the student names would have been anonymised as well, because research [suggests] how we mark [might be] based on … prejudice as well.

Some centres provided other information within cover letters, such as any extenuating circumstances. Again, it is possible that such information may have had some bearing on some moderator’s decisions (either implicitly or explicitly). However, this kind of information was noted to be a rare occurrence.

> *Moderator 1*
> We did have one where [the teacher] had been off on long-term illness for six months, [and] the stand-in teacher hadn't been in for a month… So you take that into consideration. At the end of the day, the kids have had a raw deal, and you've got to come down on their side if you can really.

For the extended project qualification, students were expected to give some explanation of why they had chosen their particular topic of study, and occasionally they gave their career ambitions here. Moderators' expectations may again have sometimes been shaped by this information.

> *EPQ Moderator*
> I suppose in your mind you're aware that if they're going to be applying for law, they're going to be probably an A student. But [I] don't get too carried with it because that doesn't necessarily mean [they will be].

Despite some moderators taking note of centre/candidate information, other moderators tried not to be led by such information.

> *Moderator 2*
> I don't, in the nicest possible way, care about who this person is. I mean, I've got their name on here but I've got their candidate number, [and] to be honest I work with candidate numbers. So they are a number to me, they are a mark to me.

4.2.2 Centre's marks, rank order, and marking trends

All moderators were aware of what mark the centre had awarded each script, before they evaluated it themselves. This clearly fed into their expectations about the likely quality of the work, and helped them know what to look for in relation to the marking criteria (eg they focussed upon the top band level descriptors when moderating a script with high centre marks). Some moderators looked at the full breakdown of centre marks at this stage (ie for each assessment objective), whereas others only looked at the total marks. After moderating the first script in the sample, moderators developed expectations for later scripts, based on the rank order of centre marks.

> *Moderator 2*
> I tend to start by looking at the mark that the centre gave it, just to give me an idea of what sort of things I should be looking for if I'm going to agree with the mark.

> *Moderator 3*
> The teacher thinks that this candidate isn't quite as good as the last one, and so I'm going to bear that in mind when I'm marking it. If the teacher's right, then I shouldn't be giving as many marks for the second candidate… as I did for the first.

After reviewing a number of scripts from the sample, several moderators began to pay attention to any trends in the centre's marking. For example, they began to recognise if the centre is marking too leniently, or too severely, in comparison with their own standards. This again fed into their expectations of later scripts in the sample, as they may expect the same to occur. However, some moderators again tried to pay less attention to such trends.

> *Moderator 5*
> I suppose my thinking is in a sense coloured by the fact that I've already marked one of them, and that's out. So the marks probably are going to be adjusted.
>
> *Moderator 7*
> Just because I've taken six off the first [script] doesn't mean then that the rest will follow, so I won't use that to sway my decision… I wouldn't think, "oh, I had to take five marks off the poetry on the last one, I need to do it on this one". I'd start from scratch, because you never know, they might have just had a bad five minutes!

After marking their students' work, centres are expected to undergo a process of internal standardisation, to ensure that the mark scheme has been applied consistently between different teachers in the centre. Some moderators looked for evidence of this (eg comments or marks made by more than one teacher on the documentation accompanying the scripts), and whether or not this evidence was found may have affected their expectations about the likely consistency of marking.

> *Moderator 2*
> The theory would say that if you've got more than one marker in a school if they've not standardised somehow or their practice isn't in place then it should lead to a problem.
>
> *Moderator 7*
> Normally if they've done some form of internal standardisation or internal moderation, they tend to be a bit more accurate than if they haven't.

4.2.3 Centre comments

Centres are also expected to give some justification for their marks. Some moderators read these comments before reading the script, whereas others read them after reading the student’s work (to avoid biasing their initial judgments). For those who read them first, comments were sometimes seen as offering a useful insight into the centre’s decision making, and again built up expectations for the quality of the work. However, comments were not always seen to be useful, depending on their content.

> *Moderator 1*
> Before I read the report, I have a look at the teacher's comments… those should help me to understand how the teacher's actually arrived at the marks for the different criteria that it's assessed.
>
> *Moderator 10*
> Sometimes I do [look at the comments first] and sometimes I don't. If I've started having doubts about the first few scripts I've looked at then I'm more likely to go straight to the teacher's comments on the next script… I'm going to think to myself, ‘right, what have they got to say about this one then?’
>
> *Moderator 5*
> The system requires you to look at the centre marks. My practice is not to read the centre comments… There’s a real danger that you’re influenced by the centre’s spin on things. And so my tendency is to read [the main body of the script] through first.

4.3 Forming first impressions

4.3.1 Surface features of the script

After reviewing the relevant documentation, moderators began to review the actual body of the script. Many moderators noticed several surface features of the work that fed into their initial impressions of script quality, which they took note of before reading any substantial part of the report.

For example, some initial impressions of quality were immediately made by noting the length of certain sections, such as the reference list or bibliography, or of the overall work. Various different conclusions were drawn based on these perceptions.

> *Moderator 3*
> Looking at that [script], I’ve weighed it, and it's about what a C [grade] candidate would write in quantity.
>
> *Moderator 4*
> There’s a source pasted onto the essay… It doesn’t mean to say it’s not a good essay, but… It’s often a sign of a candidate that’s perhaps got limited ability... I think it’s the strategy used by some students with lower literacy.
>
> *Moderator 8*
> I see a page and a quarter, closely typed. It tends to suggest there’s been some in-depth reflection rather than just a quick summary and get it out of the way.

The title of the work was also immediately noticeable, and any doubts about the appropriateness of the title led to doubts about the likelihood of the written report being of a high standard.

> *Moderator 5*
> So we’ve got a title here… And I’m going, “hmm…. So your challenge there is to define [the topic] in a way that you can effectively answer that question”. So I’m immediately going, “this is going to be difficult!”

Several moderators also commented on immediately noticeable errors such as spelling mistakes to be indicative of a lower ability candidate, again leading them to expect the rest of the assignment to be of a lower quality piece of work.

> *Moderator 1*
> Immediately you can see there's some punctuation and spelling errors, so you know that this is not going to be that high achiever really.
>
> *Moderator 3*
> The strange thing here was, I noticed it this morning, she hasn't got a capital letter for her surname. And that rings bells up there… But it’s not something that is going to sway the mark yet. I've only read her name.

During the course of this research, some moderators reviewed scripts that contained material that was inadmissible according to the assignment specification, or evidence of excessive supervision (ie beyond the level that is generally allowed). Although moderators took note of such infractions, and fed this back to the centre in their reports, this did not seem to have any impact on the mark they awarded.

> *Moderator 3*
> I got to this page on this one and I'm thinking, hang on, they're not allowed to do that… They're not allowed to do that.
>
> [Interviewer: Did that have any bearing on the marks you eventually gave?]
>
> No, because the teacher encouraged it, because all of them did it.
>
> *Moderator 8*
> [Interviewer: If the centre had given too much guidance, would that affect the marks you thought the candidate deserved?]
>
> I don't know whether it'd affect the mark; it'd certainly affect your feedback to the centre… you'd politely mention that the emphasis should be on the student; this is a student-led project.

4.4 Reading and evaluating

Moderators generally read the main body of text from beginning to end, going back to re-read certain sections once they had finished the initial read-through. This part of the process has been termed 'reading and evaluating' because these two processes occurred concurrently, with each statement being evaluated by moderators as they read it. Impressions of script quality were therefore continuously developed as moderators read through the work. However, final judgements were not made until the whole script had been reviewed (final judgements shall be discussed in the next section).

There was some variation in how thoroughly moderators read the main body of the report. Similar to that reported by Crisp (2017), some moderators tended to skim-read, or skip parts of the script, whilst others read more thoroughly; others fell somewhere in between. However, those that skim-read did note that they became more thorough when they began to have concerns about the quality of marking, or the script was inconsistent with their initial expectations/impressions of quality.

> *Moderator 5*
> If you've read the first part and [the teacher knows] what they're doing … then you can maybe move on to the conclusion. Because if you've looked at it and you think, yes, she's top band, [you] don't necessarily have to

> read the whole thing… But having said that, some of them, the difficult ones, you do end up reading every [word].
>
> *Moderator 6*
> I do [read every word], but then that's just me and [my] guilt, because otherwise I'd be thinking I'm stealing from somebody. That's why it always takes me so long to moderate.

For the extended project qualification, candidates have the option to submit an 'artefact' (e.g., an artwork or performance piece), rather than an essay based project. One moderator noted the difficulty in thoroughly reviewing certain artefacts, such as those that contain lengthy recordings of performances. However, although (s)he felt unable to watch most of the recording, (s)he was nevertheless able to review the written report that accompanied the artefact more thoroughly.

When reading/evaluating the work, moderators were able to draw upon a number of resources to help them develop their impressions of script quality, as I shall now discuss.

4.4.1 Content and quality of the script

Of course, the main factor that moderators focussed upon was the content of the written work itself. For example, the greatest number of comments made during the think-aloud exercises were related to the strength and appropriateness of the arguments being made by the student, as well as the depth of analysis and understanding being demonstrated.

> *Moderator 4*
> The most important things that we all have to focus on are the addressing the question [and] the use of the sources.
>
> *Moderator 8*
> [The student] evaluates her strengths and learning, concludes with a thoughtful and well researched discussion. So again, everything I've seen so far has made me think, yeah, this sounds pretty good.

Other aspects of writing were also commented upon, such as the overall structure of the work, any spelling/grammar errors and the effective use of subject-specific terminology. These again fed into moderators' overall evaluation of the work.

> *Moderator 1*
> Good use of specific words and phrases. Terminology's good there for the subject.
>
> *Moderator 6*
> She also structured it very well... it was broken up into different structured elements that each chapter actually reached a valid conclusion. And then she brought those conclusions together in her main conclusion to answer her question. So it was very focused.

4.4.2 Marking criteria

Evaluations of the work were usually framed around moderators' understanding of the marking criteria (whether they physically had a mark scheme to hand, or were referring to their internalised understanding of this document). Thoughts of how the work fitted into the level descriptors tended to be done on a fairly holistic and changeable basis during this stage, with final decisions on marks being left until the end. Key words found within the level descriptors often helped shape and focus moderators' thinking here.

> *Moderator 4*
> I'm now thinking, it can't be anything more than band 3. Band 4 is basically 'consistent' analysis, 'consistent' use of sources… This hasn't got that consistency, so I'm thinking it's got to be band 3, or less.
>
> *Moderator 7*
> They have got 'some' [understanding], and I'm using this word 'some'… If the descriptor is 'sound', really you're looking for that level of understanding. And so at this stage I'm thinking, well, there's only some understanding here.

4.4.3 Centre annotations

All moderators paid some attention, although not on every script, to the annotations made by the centre (eg comments made in the page margin). Where used, moderators sometimes found these annotations helpful in terms of understanding why the centre had awarded certain marks, and of highlighting aspects of the work that they may have otherwise missed. Not all annotations were equally helpful to the moderators however, and ticks in particular were noted to be generally unhelpful.

> *Moderator 1*
> As I'm reading through it, I can look at where the teacher has thought that the pupil had hit the criteria. Sometimes I'm in agreement with that; sometimes not.
>
> *Moderator 6*
> This [annotation] is really helpful because you can actually say, “oh, actually no, I don’t particularly agree with that one”… So I know [I am] out of sync with this centre and that shows me why.
>
> *Moderator 9*
> [Ticks are] not particularly helpful. It’s not always clear what they’re ticking and why they’re ticking it… It’s much more helpful… [when] I can see how they’re justifying what they’re saying… I wouldn’t say that centres that don’t put [comments] on there are disadvantaged, it just means I have to work harder, I would say.

Some moderators tried to ignore any annotations as they were reading through the script, to avoid their judgements being biased by the centre. However, their efforts to ignore them did not always appear to be successful.

> *Moderator 2*
> I tend not to read those comments, I tend just to read the essay and ignore [them] completely, partly because they can be misleading and partly because [they] can influence my thinking as I’m reading it.
>
> [When reviewing a subsequent script:]
>
> I did notice that the teacher had written “incorrect”. Now that can be helpful when I’m moderating because I might be marking a topic that I’m not an expert on entirely… And again you’ve got the teacher on the left commenting that there’s a counterargument coming, and as I read through that you can see that they’ve got these sort of counterarguments.

The degree to which moderators focussed upon any annotations seemed to depend somewhat upon how much they ‘stood out’ to them. For example, more legible and more prominently placed comments were more likely to be noticed.

> *Moderator 2*
> The teacher comments on this [script] are, because they're shorter, they sort of stand out more. And again, not that I necessarily use them, but sometimes they pop out at you.
>
> *Moderator 5*
> A lot of them are illegible… In the case of this one I haven't really picked up much on the annotations to be honest, partly because it's pencil.

4.4.4 Other scripts

When reading through the text, several moderators noted that they found it helpful to think about how the current script compared to ones that they have already evaluated (including standardisation scripts). This was perhaps especially true for those who reviewed scripts according to rank order. In effect, moderators were able to use earlier decisions as a benchmark for the quality of the current script, which allowed them to develop their impressions of quality. Nevertheless, it was rare for moderators to physically refer back to earlier scripts during the reading stage, but they did sometimes do this when making final judgments.

> *Moderator 7*
> So I'm starting to think whether it's enough, based as well by comparing it with this example piece from the standardisation, whether it actually is as good at that piece.
>
> *Moderator 8*
> It's not got subheadings like the previous one, but it is set out in to reasonably clear paragraphs. I notice as I flick through that I'm not spotting as many references [or] quotes as in the previous one.

4.4.5 Grades

Although grade boundaries can change year-on-year for NEAs, some moderators developed their impressions of the script quality by comparing the script to what they believed the standards of each grade level to be. This tended to be done on a fairly holistic basis during this stage (eg 'this feels like an A grade script'). Other moderators, however, avoided thinking in terms of grades, being more cautious about the fact that boundaries may be set in a different manner to those expected.

> *Moderator 3*
> Is this really an A candidate? OK, this going through my mind now: is this really an A candidate?
>
> *Moderator 4*
> When I very first started moderating quite a long time ago, I couldn’t do this. But through experience… within reading two or three paragraphs, you can usually tell whether a script is a B, C or A or an A*.
>
> *Moderator 2*
> I try and avoid [thinking about grade boundaries], because I don’t know what the grade boundaries are… that’s something that I think’s important to forget.

4.5 Making and evaluating decisions

As mentioned in the previous section, constant evaluations were being made as moderators read through the script. However, any evaluations were made on a fairly informal basis until they had finished reading the script, at which point more formal decisions needed to be made (ie by finalising their impressions of script quality). This decision making processes tended to be done on a somewhat cyclical basis, with decisions being made and evaluated/revised until moderators were confident in the outcome. As with the other stages of the process, several different resources were used to help make and evaluate their decision, as illustrated by the following quote:

> *Moderator 10*
> All these things are running against each other in your mind all the time, you know, and gradually something emerges like an overall judgement.

The remainder of this section has been divided into two parts, to reflect the different types of resources that moderators drew upon to help them make their decisions. ‘Benchmark’ resources (as I have termed them) were compared against the current script to help moderators determine which mark it deserved. ‘Other considerations’ were other factors that moderators needed to bear in mind when making their decisions, but did not necessarily help them to place the script within the mark scheme. This distinction should become more apparent as each section is discussed.

Benchmarks

4.5.1 Marking Criteria

When asked to name the resource that was the most important in making their decisions, most moderators named the marking criteria. However, as noted previously, some were only referring to an internalised understanding of this document, rather than a physical copy. One tactic was to use the marking criteria to confirm their initial thoughts made during the reading phase.

> *Moderator 1*
> I go through [the script] and I'm thinking, oh, you know, that's good analysis all the way through that, oh I like their evaluation there, they've used a good range of resources – but then when I come to actually do the marks, [the mark scheme] is always in front of me.
>
> *Moderator 2*
> So I get a holistic impression of the whole paper, make I suppose what would be my decision and then look at the mark scheme and confirm and then apply and just make sure that those criteria are set in it.
>
> *Moderator 9*
> The mark scheme is the biggest driver. That comes first.

Most moderators decided upon marks for individual ‘assessment objectives’ first, before summing these to arrive at the total mark for the script. A few, however, decided upon the total mark directly, deciding not to break marks down into individual assessment objectives. Part of this depended upon the subject. For example, moderators of the extended project qualification (EPQ) made their decisions on an objective-by-objective basis, due to the structure of the report.

> *Moderator 4*
> It was 32 or 33. That’s where it fitted in to that band when you brought the different criteria into account… There are three different assessment objectives. I could have done it like that, but because I know the mark scheme well, I kind of synthesise them automatically anyway.
>
> *EPQ moderator*
> By the time you’ve read the log you’ve probably got an idea of where you’re at with the AO1… then you’re looking at the bibliography and the referencing – you’ve got some idea of where you are on [AO2]. Then you’ve got to read the whole thing… to be able to get the AO3. And AO4

> you've already read what they'd done for the presentation from the log… So you are doing it in four sections.

The main approach taken by moderators was to decide upon a 'level' (or 'band') first (ie the most appropriate level descriptor), before deciding upon specific marks within that level. This is similar to the approach often taken by exam markers (eg Crisp, 2010). By doing this, moderators were able to narrow their decisions down in stages.

> *Moderator 4*
> So, what you've got to do is look at the criteria for each band. I was very confident that it was a band 3, because it was inconsistent, but [it] had some aspects that were good enough to bring it into band 3 rather than band 2. It wasn't consistently analytical enough… to get it into band 4. So it was 32 or 33 [marks].
>
> *Moderator 5*
> Which bands is it in? Has it met the minimum criteria for those bands?... Once you've got your minimum requirement then, OK, how far [into the] band can it go? How well have they done it?

Somewhat regularly, scripts did not fit neatly into these level descriptors. In such cases, moderators described how they needed to make 'best fit' judgements in order to assign a script specific marks.

> *Moderator 3*
> There are four things we're looking for in [assessment objective] B… I'm looking for four things to award one mark. So to do that, you've got to think about best fit… If someone has done three of those and made one spelling mistake, you're not going to say, "one spelling mistake, I'll knock a mark off"; you're going to say, "OK, nearly there, it's worth the full marks".
>
> *Moderator 7*
> The first bullet point is a bit stronger perhaps and the last bullet point is a bit weaker. So I felt that on balance then a mark at the top of band 3 would be appropriate. So it's a little bit of balancing out, just using judgement overall as well.

On occasion, moderators referred back to guidance that they had received from more senior moderators (eg from the standardisation meeting or via feedback from team leaders) to consolidate their understanding of the marking criteria. This guidance was sometimes used by moderators either to help them make decisions, or to evaluate the appropriateness of decisions already made.

> *Moderator 4*
> [The principal moderator] always says you need to have consistent analysis to be getting any higher than band 3.
>
> *Moderator 6*
> When we're at the moderation meetings they often say to us is this the best an 18 year[-old] could do? And looking at this I think wow yeah, this is the best that an 18-year-old can do.

4.5.2 Grade boundaries

In addition to using their knowledge of grades to help frame their thinking during reading, moderators also used this knowledge to help make their decisions at the end of the process. Grade boundaries served as a benchmark, against which the current script could be judged, to help the moderator decide upon a specific mark. This is in slight contrast to the reading phase, where grades were often used to form more general impressions about the quality of the script.

> *Moderator 1*
> The centre had given it 50, which was just below an A. I gave it 53, which was midway between an A and an A* because I found that that was quite a mature piece of work really for a 15 to 16-year-old pupil.
>
> *Moderator 5*
> What is quite helpful is the band criteria. If you're not sure, is it A*?. Re-read the A* criteria. And that's helpful… You're not marking to the grades, because you're marking by [assessment objectives]. But at the end of the day if… the [assessment objectives] add up to 45 it should meet the A* criteria.

Grade boundaries were also used to help evaluate any decisions that may have been made. Specifically, decisions/marks were sometimes adjusted when a moderator's marks did not align with their understanding of each grade level.

> *Moderator 3*
> If we say the C is a 38 and I'd marked one and it had been 33, I would have thought, “hang on, it reads like a C. Let's go back and see where I missed something”. And if it feels like the C, then I will find those marks.
>
> *Moderator 10*
> I'm also saying to myself, “I think 40 is probably going to be a bottom A, is that about right?”… I would find this impossible to do if I wasn't able to have some kind of idea about roughly the kind of grades that the total marks [align with].

Not all moderators used their understanding of grade boundaries to help make and evaluate decisions, however, and some acknowledged the risks associated with assuming that grade boundaries will remain stationary over time.

> *Moderator 4*
> You’ve got to be very careful there, because a C grade might differ slightly in different years. It’s actually quite consistently moderated and marked, so a C is usually between 30 and 34 for this. But it might be 29 and 33 in one year or 32 and 35, so obviously you can’t do it that way. So what you’ve got to do is look at the criteria for each band.

4.5.3 Other scripts

Previously evaluated scripts (either in the same sample or at the standardisation meeting), also served as useful benchmarks against which to make final decisions. Moderators often made comparisons between these scripts and the script currently under review (either mentally or physically). It was rare for moderators to refer back to scripts from earlier samples (at least not explicitly); comparisons were generally made with scripts from the same sample as the script currently under review, and standardisation scripts.

> *Moderator 1*
> I think originally I was up to 39 [marks]. I've gone to 37 now… after I looked at another one with the same mark.
>
> *Moderator 2*
> In the standardised scripts, I had one that we’d agreed was 14 and one that we’d agreed was 17. And I felt that [the current script] matched up with the features of the 17 one.

Again, as well as helping them to make decisions, comparisons with other scripts seemed to be a useful way of evaluating those decisions, especially in cases where moderators were doubting any decisions made (and required further clarification).

> *Moderator 1*
> What I tend to do, once I've moderated the first one, I keep going back and I think, “oh I'm not really sure on this, well what did I give that one? Is this one better or worse than that one?”
>
> *Moderator 9*
> I thought this one was a 19. So I went to these [standardisation] scripts that I’ve got here, and… that made me think, “OK,… if that one’s a 20, that one’s a 19”.

4.5.4 Intuition/experience

More experienced moderators use their internalised standards as a benchmark for decision making (all of those within the current sample had at least 5 years’ experience). Two moderators in particular noted that the most important driver of their decision making was their intuition and/or their experience. Most other moderators also acknowledged the importance of this in their work, noting that they had found the process more difficult when they had first started in the role.

> *Moderator 9*
> I can probably tell you what [the level descriptors] are without looking at them, because I’ve done it for so long… I think the first time you do it it’s really hard.
>
> *Moderator 10*
> I've got pretty clear ideas that that is not 47! I kind of know what a 47 looks like and it's a much higher quality of work than this.
> If I was to be absolutely honest, it's my experience [that is the most important resource]… It's not that I'm ignoring the marks scheme, but it's so internalised that it just informs everything I look at when I'm reading this stuff through. But for an inexperienced moderator it would be probably very different.

Other considerations

4.5.5 Centre’s marks and tolerance thresholds

As well as allowing moderators to know what to look for while reading the script, the marks awarded by the centre often helped moderators in their decision making at the end. However, centre marks were usually used by moderators to check whether their initial decisions made sense, rather than helping them to make those decisions in the first place. Although some moderators had chosen not to look at the breakdown of the centre’s marks before making their own decision (ie the mark for each assessment objective), all moderators were always aware of the total marks that had been awarded. Those that did not look at the breakdown of marks to start with, often used this breakdown to evaluate their decisions after they had been made.

> *Moderator 1*
> Initially I think [I was] up to about 39 as opposed to 33, which the centre had given… I [then] went back through it first of all and started to make some alterations, tried to see where I was in agreement. Could I actually come down with the teacher? No, it wasn't actually clear that I could.
>
> *Moderator 3*
> I've got the teacher's mark and my mark… and I'm thinking hang on, there's two differences there… what I'll do is go back and say, “well, perhaps the teacher's right”.

The default position that many moderators took was that the centre’s marks were correct, unless evidence could be found to suggest otherwise. The centre’s marks were therefore an important consideration in moderators’ decision making, and in their evaluation of their decisions.

> *Moderator 2*
> I suppose ultimately what I’m doing is I’m saying, “is the score that the school submitted appropriate?” And perhaps it could just be a yes or no that I’m giving. ‘Yes’ being a quite a broad yes, because of tolerance of plus or minus three.
>
> *Moderator 3*
> That's my starting off point in all this moderation. The teacher's right, unless I can find a reason why the teacher's wrong. I always go back to that idiom: the teacher's right. The teacher knows these kids. I don't. But that's why I'm the moderator – I'm independent… If I can [agree], I will, but if I can’t, then so be it.

> *Moderator 4*
> I don't want to change marks… You know, when it all boils down to it, you hope that they are [correct], don't you? My job isn't to change marks, only if I have to.

Some differences between the moderation of June and November assessment series were observed here, although a clear direction of effect was not apparent. One moderator suggested that (s)he was more likely to agree with a centre's marks in June due to time limitations (more scripts are usually submitted for moderation in June compared to November). Another moderator implied that (s)he was more likely to agree with centres in November, because those that submit in November are usually the more experienced, and therefore less problematic centres.

> *November moderator 1*
> [In] June sometimes we're a bit more pushed for time and you might tend to go, "well, if it's not going to alter the centre marks, I'll agree with the centre". But at the moment I'm not doing that.
>
> *November moderator 2*
> Centres who have chosen to enter in November… do tend to be less problematic centres and smaller centres… It's less likely to be a new centre in November. And the new centres can be the ones where problems could arise.

When moderators were in agreement with the centre's marks, decisions were often made quite quickly (ie minimal evaluation of those decisions was made). However, for many moderators, any disagreements with the centre (especially when beyond tolerance) made them evaluate their decisions much more carefully. Some exam boards allow one script from each centre to be deemed an anomaly, therefore allowing moderators to ignore one script that appears to have been marked out of tolerance.

> *Moderator 1*
> If my mark is way out from the teacher's marks, then I go back and I look at their annotations… to see why they've given it the mark they have, and then whether I still agree with that… If it's within tolerance, that's fine, I can live with it. If it's outside tolerance, then it needs to be looked at a bit more closely really.

> *Moderator 3*
> I've changed the mark to put it in tolerance. Because my mark for that was 13 and the teacher had given it eight… And I had another look at it and thought yeah, perhaps you're falling asleep there, 8 is nearer it.

Some moderators appeared to be much less concerned (explicitly, at least) about tolerance thresholds. The degree of this concern may possibly be related to levels of experience, as explained by the principal moderator in the following quote (the suggestion is that lesser experienced moderators tend be more concerned about this). However, it is unlikely to be totally dependent upon experience, as some of the experienced moderators in our sample did base their decisions to some degree on tolerance thresholds.

> *Moderator 5*
> If you can put a centre out of tolerance you're actually probably doing them a favour, because you're marking more stuff and then they get the information.
>
> *Principal moderator*
> [Less experienced moderators] normally make the changes but they don't make big enough ones. So they're not reluctant to put down a different mark. What they're reluctant to do I think is to make an adjustment of 10 instead of an adjustment of four… it can be seen to be safer to fudge it a bit and end up in the middle rather than put down what you really think. So it takes a hell of a nerve to turn around and say that's minus 10 on that. It takes a lot of confidence.

4.5.6 Rank order of marks

In cases where moderators were in disagreement with the centre, another consideration was whether this would change the rank order of the centre's marks (ie the order in which they had ranked the performances of their students). This was a greater consideration for some moderators than others. Some made efforts to try, wherever possible (ie where marking errors were not extreme), to resolve any rank order issues that were caused by their mark changes, whereas other did not.

> *Moderator 2*
> I'm looking for the rank order, and whether I agree with the rank order or not, because that's the most important thing. If I disagree with rank order

> then we would have to get the school to do a remark before moderation could take place again.
>
> *Moderator 3*
> If it works out that they've got the wrong rank order, the world as we know it explodes because things have to happen... So what I'm going to do is I'll look at the teacher marks after I've done the full [read-through], look at the teacher marks and think, “well, yeah, I can see why they gave that mark now”. And so my impression was wrong.
>
> *Moderator 7*
> I don’t really [worry about the rank order]. I don’t know if this comes across quite cold, but… well, I just think about the mark scheme and the comparison activity and think, well, what’s just and what mark can I justify?

The fact that some moderators appeared to be more concerned about the rank order than others may be due to differences in the policies set by different exam boards or principal moderators. This is perhaps reflected in the following quote, as this moderator has changed his/her strategy in response to changes in the guidance given to them.

> *Moderator 9*
> A long time ago… I think the expectation was, yes, you would try and avoid [changing the rank order] if you could. If you couldn’t then you’d change it… [but now] the advice to us is different.

Of those who did try to maintain the rank order (where such a decision could be justified), one of two approaches was taken. Firstly, if the changes would be within the tolerance threshold, then they were seen as unnecessary and so the moderator might revert back to the centre’s original marks. Secondly, two scripts that deserved different marks (as perceived by the moderator) could be awarded the same mark, so as to maintain the rank order.

> *Moderator 10*
> If there was a candidate say on 30 and another one on 31 and I thought it was the other way round, I wouldn't bother, because it's neither here nor there in a way.

> *Moderator 2*
> So the 33 marker which was in the additional sample I wanted to leave at 33, and I did. And then the 35 marker… [that] I did want to move to 31, I then brought back up to 33 so it sits in the same place. So, theoretically I'm now saying that these two are the same, even though I still know they're not. I still want the 35'er to be a lower but… we're talking one mark, two marks, it's negligible.

Although decisions were sometimes revised based on a consideration of rank order, moderators agreed that it was quite rare for a centre to get the rank order wrong, and so this rarely became an issue. However, because there was some acknowledgement that by moderating the sample according to the rank order of marks, moderators may perhaps be biased towards maintaining that order. The use of the aforementioned methods may also mean that moderators perceived rank order issues to be rare, because they were generally quite quickly resolved.

> *Moderator 4*
> I've been doing this for about 10 years and I think twice I've had the rank order wrong. It's rare. It's very rare.
>
> *Moderator 3*
> When you put them in rank order, you are certainly guided towards keeping them in rank order. Mentally, you say "right, this is the rank order the teacher's put; let's do the same".

4.5.7 Centre annotations/comments

Annotations and comments made by the centre were another resource that shaped moderators' final impressions of script quality, and therefore final decision making. When moderators believed that they disagreed with the centre (particularly when this disagreement was beyond tolerance), efforts were often made to understand the cause of such differences, so that they could justify the difference in marks. Centre comments/annotations served as a useful insight into the reasons for these discrepancies. As noted previously, some moderators chose not to read comments before reading the script. However, nearly all moderators noted finding these comments useful when evaluating their decisions.

> *Moderator 1*
> What I try to do is to look at everything blind. And then, if my mark is way out from the teacher's marks, then I go back and I look at their

> annotations… I try to see why they've given it the mark they have and then whether I agree with that, or do I want to keep to my original mark.
>
> *Moderator 10*
> The comments usually are very revealing in that they're not awarding the marks for the right things or they think that what the candidate is doing is absolutely 100% relevant when it isn't. So the comments quite often are very, very helpful indeed and they begin to explain maybe why the teacher has awarded marks that you don't think appropriate.

4.5.8 Fairness

The final consideration is one of fairness. When making and evaluating decisions, some moderators described efforts made to try and be fair to the candidate, or to give benefit of the doubt when struggling to decide between two marks. Fairness was especially important for some moderators when they believed that certain decisions would penalise students for reasons beyond their control.

> *Moderator 1*
> It's a child's future and so I do keep going back and thinking, "well, if the teacher's come down favourably with them, and I can justify that, that's fine".
>
> *Moderator 3*
> If they're given too much [supervision], the teacher's wrong. But then I'm thinking, "hang on, we've got 16-year-olds here. I can't penalise a 16-year-old because the teacher got it wrong".

Nevertheless, making decisions based on fairness did not always mean being generous towards one student, but sometimes meant being fair to other students in the cohort (a similar finding was also reported by Crisp, 2016). On occasion this may mean that individual candidates receive a potentially unfair mark (as perceived by the moderator), to avoid the whole centre being subjected to an unfair mark adjustment.

> *Moderator 6*
> I think ['fairness'] is quite a broad one because, if you don't mind me saying, because [fairness] is actually fairness to the student or fairness to other candidates who are submitting.

> *Moderator 4*
> So, if you've got one that doesn't match all the others, it's unfortunate that that candidate, they might gain, they might not. It depends which way it actually has been marked. But basically you've got to do the fairest possible thing that you can do. So sometimes you take that option of ignoring one of them [(ie declaring that script an 'outlier')] to hopefully reflect fairness for all the rest of the students… You know, you will get the occasional one that doesn't fit in with all the rest. Perhaps they weren't concentrating when they marked it or missed something.

4.6 Provisional and final decisions

All moderators stated that once the above process of making and evaluating decisions had been completed for individual scripts, decisions were provisional until the rest of the sample had been moderated. This was due to a recognition that new information could be gained from later scripts, and to allow them to bear in mind certain considerations across the whole sample, such as maintaining the rank order across the range of scripts.

> *Moderator 1*
> No, that [decision is] preliminary… Because I might look at the next one down and think, "oh, actually the rank orders aren't right there, so I do need to go back and double check again".
>
> *Moderator 4*
> It's too early at that stage. You need to see some more scripts… It's not a final decision, because I have to take into account the rank order.

After finalising their decisions for the sample, and after reviewing any further samples that might be needed, moderators were required to write a short report outlining justifications for their decisions, and to provide feedback to the centre. This report was typically based upon notes that had been made throughout the process. Feedback to the centre was generally seen as a key part of the process, to help the centre improve the quality of their marking for the next assessment series. On occasion, moderators used this feedback as one last evaluation of their decisions.

> *Moderator 5*
> Sometimes you're writing feedback and think, "well, actually?" So the feedback process is a double check. I have changed stuff at feedback.

After submitting this feedback and their report to the exam board, moderators moved on to the next sample that they had been allocated.

5 Discussion

The purpose of this research was to develop a greater understanding of how moderators make their decisions, with a particular focus upon moderators' use of the various mental and physical resources available to them. By exploring these processes during 'live' moderation activities, a new model of the decision making process has been developed. Although differences were noted in moderators' use of resources, findings did suggest overall consistency in the series of steps taken by moderators (see Figure 1). As well as offering us further insight into current approaches to moderation, these findings can be used as a foundation for discussions around what might constitute best practice, therefore also offering us insight into how the moderation of NEAs in England can be improved.

Many aspects of the judgemental aspect of the moderation process do support the validity of the process. First, the fact that moderators took largely the same series of steps demonstrates overall consistency in the system. This consistency was apparent both within the same subject area/exam board, and across different subjects/exam boards. Second, as one would hope, the use of the marking criteria was central to moderators' decision making and end-of-process evaluations. Third, statements made during the think-aloud aspects of the interviews showed that moderators were all focussing on appropriate features when reading the work (ie the strength of the argument/depth of understanding etc.). Fourth, the sheer range of resources that moderators used to help them develop and finalise judgements also highlights the general thoroughness of their work. Fifth, the standardisation meeting was accepted as a key process by moderators, and they appeared to take on board the guidance given to them by more senior members of the moderation team. Finally, feedback to the centre was seen as an important part of the process by the moderators that took part in this research, thus helping centres to make any necessary improvements to the consistency of their marking.

However, contrary to Crisp (2017, p. 16), who reported "[no] threats to validity in relation to moderator judgements" and "no evidence of bias in judgements", some of the current findings may suggest the potential for bias, from which potential threats to validity could arise. For example, several aspects of the current findings suggest risks of confirmation biases in moderators' judgments (ie a tendency towards agreeing with a centre's marks). Several moderators stated that the centre's marks were a starting point for all decisions (this was also reported by Crisp, 2016), assumed correct unless proven otherwise, which strongly suggests an 'anchor-and-adjustment' approach to moderation. Importantly, research has shown that such approaches can lead to increased levels of agreement between two markers compared to when the original marks are not known, which has been attributed to a tendency towards making conservative adjustments from the anchor (ie the centre's marks) (Garry, McCool, & O'Neill, 2005). Other research has shown that initial dispositions (based upon what I have termed expectations and first impressions) can affect one's interpretation of later information (eg that gained while reading the main body of a script): greater attention and weight is often given to disposition-consistent information than disposition-inconsistent information during impression development and decision making (see Bond, Carlson, Meloy, Russo, & Tanner, 2007). Given that expectations are primarily based upon information provided by the centre, moderators' judgments may again tend towards agreement with the centres' marks. The apparent hesitation of some moderators to award marks out of tolerance and to alter the rank order of centres' marks again suggests that small mark adjustments may not have been made when perhaps they should have been.

The fact that scripts are not anonymised with regards to centre and candidate names may also create possible threats to the validity of moderators' judgments. For example, it was implied (and explicitly stated in some cases) that moderators may have been influenced to some degree by the reputation certain centres held. Knowledge of the student's name may also give rise to bias, as studies have shown that student's demographic characteristics (some of which may have been indicated from their name) can bias marker decision making (eg see Brooks, 2012; Fleming, 1999; Harlen, 2004). There is no reason to assume that moderators would not also succumb to the same effects. Students' career ambitions were also stated on documentation for the extended project qualification, again affecting moderators' beliefs about the work in some instances. These points suggest that the process may be improved by anonymising details of centres and students.

Further suggestions for improvement might be made via a consideration of the different approaches to moderation. For example, some moderators worked through each sample according to the rank order of centre marks; others did not. A positive aspect of the former approach may be that moderators are perhaps better able to use earlier scripts as benchmarks for later scripts, but the latter approach is perhaps less prone to a desire to maintain the rank order. Further work may be needed to determine which (if either) approach is the most desirable. The use of grade boundaries as a benchmark for thinking/decision making might need particular scrutiny, as some moderators did rely quite heavily upon their knowledge of grade boundaries, despite the fact that these are subject to change. This could therefore prove problematic should those boundaries change unexpectedly during awarding. Finally, the reliance on internalised understandings of the mark scheme by some moderators may also need to be addressed. As was discussed by Brooks (2012), some research has suggested that the use of internalised understandings of mark schemes can lead to a shift in standards, due to the introduction of the examiner's personal beliefs and expectations into their evaluations of students' work (also see Bloxham, 2009).

As this work was largely exploratory in nature, there are several potential avenues for further investigation that might address some of the limitations of this study.

Further work is needed to confirm the generalisability of these findings. Though the selected approach allowed for a deep exploration of the processes available to moderators, it was not possible to strongly demonstrate any generalisable differences between subjects, exam boards, or levels of study (although some possibilities have been raised). Further work is therefore needed to confirm whether the findings reported here apply across a range of different contexts, or where differences may exist (and why). Future work might also investigate whether there are any differences between moderators of different levels of experience, as all those included within the current study were all fairly well experienced.

Researchers might also wish to explore the reasons why moderators use certain resources, while other moderators do not. Differences in approaches to moderation might result from differences in standardisation or feedback practices between different moderation teams, and so it may be useful to look more closely at these areas. Similarly, moderators appeared to use some resources to a greater degree for some scripts compared to others. It may be interesting to investigate what features of scripts determine the approach to moderation that is taken.

Finally, although potential sources of bias have been identified in the preceding discussions, the degree of impact (if any) that these biases may have on the validity of moderators' judgements needs to be determined. Further research is ultimately needed to determine whether these differences do indeed pose threats to validity, or whether different approaches are simply means to achieve the same (valid) ends.

> *Moderator 10*
> I would hope that although [we take slightly different routes], we probably are pretty close to where we end up. I mean, I don't think there can be one standard way of doing this coursework moderation because there are so many different things. How you prioritise all those, and handle it personally, I think is bound to differ.

References

AQA. (2013). *Moderation of internal assessments*. Retrieved from http://www.aqa.org.uk/exams-administration/coursework-and-controlled-assessment/moderation

Bloxham, S. (2009). Marking and moderation in the UK: False assumptions and wasted resources. *Assessment & Evaluation in Higher Education*, *34*, 209–220. http://doi.org/10.1080/02602930801955978

Bond, S. D., Carlson, K. A., Meloy, M. G., Russo, J. E., & Tanner, R. J. (2007). Information distortion in the evaluation of a single option. *Organizational Behavior and Human Decision Processes*, *102*, 240–254. http://doi.org/10.1016/j.obhdp.2006.04.009

Brooks, V. (2012). Marking as judgment. *Research Papers in Education*, *27*, 63–80. http://doi.org/10.1080/02671520903331008

Crisp, V. (2010). Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education*, *36*, 1–21. http://doi.org/10.1080/03054980903454181

Crisp, V. (2017). The judgement processes involved in the moderation of teacher-assessed projects. *Oxford Review of Education*, *43*, 19–37. http://doi.org/10.1080/03054985.2016.1232245

Daly, A., Billington, L., Chamberlain, S., Meyer, L., Stringer, N., Taylor, M., & Tremain, K. (2011). *Principles of moderation of internal assessment*. Manchester, UK: Centre for Education Research and Policy. Retrieved from https://cerp.aqa.org.uk/research-library/principles-moderation-internal-assessment

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: The MIT Press.

Fleming, N. D. (1999). Biases in marking students’ written work: Quality? In S. Brown & A. Glasner (Eds.), *Assessment matters in higher education*. Philadelphia, PA: SRHE and Open University Press.

Garry, J., McCool, M. A., & O’Neill, S. (2005). Are moderators moderate?: Testing the “Anchoring and Adjustment” hypothesis in the context of marking politics exams. *Politics*, *25*, 191–200. http://doi.org/10.1111/j.1467-9256.2005.00243.x

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second langauge research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Harlen, W. (2004). *A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes*. London, UK: EPPI-Centre. Retrieved from http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=116

JCQ. (2016). *Legacy GCE unitised AS and A-level qualifications; ELC and Project qualifications: Instructions for conducting coursework*. London, UK: Joint Council for Qualifications. Retrieved from http://www.jcq.org.uk/exams-office/coursework/instructions-for-conducting-coursework-2016-2017

Johnson, S. (2011). *A focus on teacher assessment reliability in GCSE and GCE*. (Ofqual Report 11/4807). Coventry, UK: Ofqual. Retrieved from https://www.gov.uk/government/publications/a-focus-on-teacher-assessment-reliability-in-gcses-and-a-levels

OCR. (2015). *Admin Guide: 14-19 Qualifications*. Retrieved from http://www.ocr.org.uk/Images/252669-admin-guide-and-entry-codes-14-19-qualifications-2015-16.pdf

Ofqual. (2011). *GCSE, GCE, principal learning and project code of practice*. Coventry, UK: Office of Qualifications and Examinations Regulation. Retrieved from https://www.gov.uk/government/publications/gcse-gce-principal-learning-and-project-code-of-practice

Ofqual. (2013). *Review of controlled assessment in GCSEs*. (Ofqual Report 13/5291). Coventry, UK: Office of Qualifications and Examinations Regulation. Retrieved from https://www.gov.uk/government/publications/review-of-controlled-assessment-in-gcses

Ofqual. (2015). *GCSE (graded A* to G) controlled assessment regulations*. (Ofqual Report 15/5742). Coventry: Ofqual. Retrieved from https://www.gov.uk/government/publications/gcse-controlled-assessment-regulations

Pearson. (n.d.). *Coursework moderation and mark adjustments (an explanation for centres)*. Retrieved from http://qualifications.pearson.com/en/support/support-topics/results-certification/moderator-reports.html

Strauss, A., & Corbin, J. (1994). Grounded theory methodology: An overview. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research* (pp. 273–285). Thousand Oaks, CA: Sage Publications.

Williamson, J. (2016). Statistical moderation of school-based assessment in GCSEs. *Research Matters: A Cambridge Assessment Publication*, *22*, 30–36.

Wilmut, J., & Tuson, J. (2005). *Statistical moderation of teacher assessments*. London, UK: Qualifications and Curriculum Authority. Retrieved from http://oucea.education.ox.ac.uk/wordpress/wp-content/uploads/2011/06/Wilmut-2004-review-of-TA.pdf

WJEC. (2015). *Internal Assessment Manual: June 2016 Series*. Retrieved from http://www.wjec.co.uk/exam-officers/related-documents.html?category=internalAssessmentSubmissionOfMarks

Appendix – Details of the methodology

Design

In order to gain a comprehensive insight into the usual processes employed by moderators, a mixture of ‘think-aloud’ and more traditional one-to-one interviewing methods were employed. In think-aloud methods, participants are asked to perform an activity while verbalising everything that they are thinking, looking at, and/or doing. Such methods allow one to access mental processes that would be otherwise unavailable via other means (eg traditional interviewing or behavioural observation) (Gass & Mackey, 2000). Although there are some limitations associated with this method (eg see Brooks, 2012), it is able to offer a relatively more accurate depiction of underlying cognitive processes than methods such as traditional interviewing (see Ericsson & Simon, 1993), because participants’ responses are mostly unaffected by any leads or cues given by the researcher. However, as details offered by think-aloud methods are dependent upon what is offered by the participant, follow-up interviewing can be helpful to clarify any comments made during the think-aloud exercise, or to explore any factors that were not mentioned during think-aloud. A balance has to be struck, however, between the comprehensiveness of this combined approach, and the effects of hindsight introduced by follow-up questioning.

Think-aloud methods are best done during live processes, rather than retrospectively (Ericsson & Simon, 1993). However, research on actual ‘live’ moderation activities was not possible in this instance, to avoid disrupting exam boards’ usual processes and remove the risk of affecting outcomes for students. As such, a delayed recall approach was necessary. Given that the richness and accuracy of recall is time sensitive (see Ericsson & Simon, 1993; Gass & Mackey, 2000), moderators were interviewed as soon as possible after they had completed their live moderation (this is discussed in the following ‘procedure’ section).

Thematic analysis was chosen as the analytical approach for this research. I also employed a grounded theory approach to achieve a richer, more comprehensive model of the moderation process. Grounded theory is a method of theory development which is grounded in qualitative data (for an overview, see Strauss & Corbin, 1994). A key feature of this method is the cyclical process of data collection and data analysis, in which analysis drives further iterations of data collection; the aim being to verify or further explore hypotheses made during initial waves of collection/analysis (Strauss & Corbin, 1994). Following this approach, a first wave of data collection was carried out in June 2016. Findings were analysed and a second wave of data collection was conducted in November 2016, with the interview schedule being adapted to verify and further explore the outcomes of Wave 1.

Usually, the goal of this cyclical process would be to achieve ‘theoretical saturation’ (ie when new data analysis yields no new information to develop the theory). However, as the moderation of GCSE/A level NEAs only happens twice a year, it was decided to limit the current project to 2 waves of data collection.

Recruitment

Ten moderators in total were recruited from 4 exam boards (EBs). Of these, 2 were moderators of GCSE business studies (EB1), 3 were moderators of GCSE history (EB2), 2 were moderators of GCSE English (EB3), and 3 were moderators of a level 3 extended project qualification (EB4). Note that this design does not allow for meaningful comparisons between different subjects/EBs, but rather was driven by availability and was intended to capture a range of experiences from different contexts and backgrounds. The subjects studied were selected by the exam boards, with this decision being largely based upon availability and suitability for the project. All participants had at least 5 years’ experience of moderating for their unit. Each participant was paid £200 plus travel expenses in exchange for their time.

Procedure

After having the purpose of the study explained to them, and providing consent to take part, participants were left alone for approximately two hours to conduct their usual moderation activities uninterrupted. This was done to avoid affecting their live moderation judgments. Participants were instructed to try and complete moderation in full for at least one centre, although some were able to moderate all of two centres’ sub-samples during this time. Others were only able to work through 3 or 4 scripts. Depending on time, there was a short lunch-break for some participants after completing their moderation, whilst others progressed straight to the interview stage. Any gap in time between stages was kept to a minimum (usually under half an hour).

For each script in turn (in the same order as before), participants were asked to think aloud while repeating the same processes as earlier in the day. To avoid researcher interference, this stage was largely unstructured and uninterrupted; participants were left to freely declare their mental processes. Once they had finished, they were asked a series of questions to clarify or follow-up on comments made or aspects of their decision making. This process was repeated on a script-by-script basis. Due to time constraints, not all scripts within a sample were reviewed in this manner. Rather, we aimed to work through the first few scripts of the sample, and then any that stood out as being more interesting cases (eg to explore where there had been some disagreements with the centre). The final number of scripts that were reviewed in full was ultimately dependent upon each moderator and their sample, but at least 3 scripts were discussed for each participant. Once enough scripts had been reviewed, a number of questions were asked about the sample overall, along with questions about their moderation activities more generally.

For the first wave of data collection (June), the mean length of the think aloud/interview stage across the sample was 82 minutes, with a range of 54 to 103 minutes. For the second wave (November), the mean length was 116 minutes, with a range of 92 to 142 minutes. This increase in duration between sessions reflects the additional interview questions included as part of the grounded theory approach.

Audio recordings were transcribed by an external transcription company and a sample of these transcripts were checked by the researcher for accuracy. Transcripts were coded and analysed (using thematic analysis) using ‘NVivo 10’ software for Windows.

We wish to make our publications widely accessible. Please contact us at publications@ofqual.gov.uk if you have any specific accessibility requirements.

This publication is available at www.gov.uk/ofqual.

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346