Insights into teacher moderation of marks on high-stakes non-examined assessments

Victoria Crisp Research Division

Introduction

Where teachers assess their students' work for high-stakes purposes, their judgements are standardised through professional discussions with their colleagues – a process often known as *internal moderation*. This process is important to the reliability of results as any inconsistencies in the marking standards applied by different teachers within a school department can be problematic.

This research explored internal moderation practice for school-based work contributing to high-stakes assessments in England, Wales and Northern Ireland, with a focus on General Certificate of Secondary Education (GCSEs). Since their introduction in the 1980s, GCSEs in some subjects have involved a component of work that students conduct in the classroom rather than in the exam room. The nature of this work and the restrictions around it (e.g., how much time is allowed, whether it can be partly conducted at home) have varied and the number of subjects with a non-examined element has reduced over time. Broadly speaking, the work tends to involve some kind of project or extended piece of work that could not, realistically, be conducted within the time and limitations of an exam situation. Generally, such work is marked by the students' teachers and externally moderated by examiners appointed and trained by the awarding organisation (AO). The procedure for *external moderation* involves the school submitting a list of their students' marks and the 'moderator' selecting, in line with the AO's guidance, a relatively small number of students' work to review. Their selection ensures that students across the ability range are sampled. The moderator reviews the teachers' marking of the student work and determines whether the marking at the school is in line with national standards or requires adjustment (see Crisp, 2017, for insights into the processes involved in external moderation).

Where there is more than one teacher marking the student work in a particular school for a particular qualification and subject, there is a requirement for marks to be internally moderated before submission to the AO (i.e., before external moderation is conducted). Internal moderation can involve one teacher (e.g., the head of department) evaluating the marking of the other teachers against their own, and adjusting marks if needed, or a group meeting of the teachers within a school department where they compare and discuss how they would each mark the same example pieces of student coursework, agree on the marks for these, and subsequently adjust the marks given to other students if needed. The aim is for a school's set of coursework marks across all classes for a particular subject to have been marked to the same standard such that students are placed in an accurate rank order in respect of the specified criteria. Some General Certificate of Education Advanced levels (GCE A levels) also involve a non-examination element and tend to be assessed following a similar model to that for GCSEs.

Where internal moderation is a group activity, the process could be similar to that of consensus moderation described by Sadler in the context of higher education:

> *Consensus moderation starts with a sample of student responses drawn from the course pool. Working independently, all assessors mark all responses in the sample. For each, they record their provisional judgement and their reasons for it. Markers then convene as a group, individually present their decisions and rationales, and deliberate them until consensus is reached.*
>
> (Sadler, 2013, pp.7–8)

The approach involves comparison and alignment of judgements of student work against stated criteria, leading to clarification of interpretations of criteria and the development of shared understandings.

A number of studies have usefully explored the judgements involved when teachers assess student work. For example, Cooksey, Freebody and Wyatt-Smith's (2007) detailed coding and analysis of the influences on teachers' assessments revealed the complexity of their judgements and the varying strategies of different teachers even when applying national benchmark criteria. The knowledge and skills that teachers need have also been discussed. Drawing on Sadler (1998), Klenowski and Wyatt-Smith (2013) proposed that teachers making judgements about student work need to be able to utilise: knowledge of the content to be assessed; deep knowledge of the assessment criteria; and evaluative skills developed from previous experience of judging student work on similar tasks. A newly qualified teacher should already have the first of these, but the other two elements need to be developed through experience. Teachers will be provided with a set of written marking criteria to use but these could be subject to differing interpretations if applied by isolated individuals (Johnson, 2013). The provision of written exemplars may help with understanding the intended meaning of criteria (Sadler, 1998) but involvement in consensus moderation is also likely to be valuable to understanding criteria (Johnson, 2013), and is recognised to be a useful professional learning opportunity for teachers, building their assessment capacity (Harlen, 2005; Klenowski & Wyatt-Smith, 2010; Smaill, 2012). Common interpretations of criteria are developed through discussion of evidence depicting the qualities represented in the criteria (Klenowski & Wyatt-Smith, 2013). In the context of school-based assessment in Queensland, Australia, Klenowski and Wyatt-Smith (2010) argued that moderation meetings provide:

> *… an opportunity to generate new knowledge and new ways of knowing as teachers draw on their individual tacit and individual explicit knowledge and the group's tacit and explicit knowledge, and use this knowledge as a tool of knowing within a situated interaction with the social and physical world*
>
> (Klenowski & Wyatt-Smith, 2010, p.121)

Various examples are given, such as being able to check that similar skills are taught and assessed, increased confidence in the understanding of achievement expected at particular levels, and a shift from individual practice to shared practice and improvement (Klenowski & Wyatt-Smith, 2010). Klenowski and Wyatt-Smith (2013) related this process to Cook and Brown's (1999) notion of 'bridging epistemologies' in which individuals' tacit and explicit knowledge (that would bear on the individual's judgements) are revealed, and ways of knowing are generated through the group process of working together to articulate their understandings of criteria and develop a shared perspective. This is not dissimilar to Wenger's (1998) theory of communities of practice and shared understandings and practices developing through participation.

The role of collaboration amongst teachers has been explored by Allal and Mottier Lopez (2014). They drew on Cobb, Gravemeijer, Yackel, McClain, and Whitenack's (1997) notion that human judgement involves a reflexive relationship between an individual's psychological processes and shared social practices. Within this view, it is argued that meaning is not identical in the minds of all those involved but that interactions between participants allow 'taken-as-shared' meaning to emerge which guides activity (Allal & Mottier Lopez, 2014). Evidence suggests that collaborative assessment activities can facilitate this process of 'deprivatisation' and the construction of shared practices (Allal & Mottier Lopez, 2014; Black, Harrison, Hodgen, Marshall, & Serret, 2011). Such theories and evidence suggest that it is likely to be important that teachers have opportunities to be involved in collaborative assessment activities.

Perhaps the most directly-relevant research to the current study is Wyatt-Smith, Klenowski and Gunn's (2010) analysis of recorded teacher talk during consensus moderation meetings of teachers in the Queensland context. Their research identified that teachers move back and forth between:

> *(1) supplied textual artefacts, including stated standards and samples of student responses;*
>
> *(2) tacit knowledge of different types, drawing into the moderation; and*
>
> *(3) social processes of dialogue and negotiation.*
>
> (Wyatt-Smith, et al., 2010, p.59)

They concluded that the written assessment criteria are 'insufficient to account for how the teachers ascribe value and award a grade to student work in moderation' (p.59), and emphasised the social and cognitive elements of moderation practice. A tension was found between criteria that teachers carry 'in their head', developed through experience, and the stated criteria. The former was influential in judging ability but essentially unstated, though assumed to be common with those held by others. Wyatt-Smith et al. (2010) concluded that this tension is not necessarily a sign that teacher judgement is flawed or biased, but that assessment judgement involves a number of challenges.

Internal moderation procedures aim to ensure the consistency with which marking standards are applied within a school, both in terms of the reliability of teacher judgements and standards over time (Klenowski & Wyatt-Smith, 2013). In the context of non-examination elements of GCSEs and A levels, without consistency of marking within a school in terms of establishing an appropriate rank order, external moderation procedures would be difficult to implement appropriately. Aiding teacher development and improving the accuracy of future marking are likely to be additional aims of internal moderation. This study sought to improve our understanding of internal moderation practice in the context of GCSEs and A levels.

Method

This research involved the use of three complementary methods: semi-structured interviews; mock internal moderation sessions; and a questionnaire survey.

The interviews and moderation sessions were conducted with GCSE teachers with experience of internal moderation of coursework for English/English Literature, Geography, or Information and Communications Technology (ICT). These subject areas were selected to represent a variety of types of student work. The marking criteria for GCSE coursework in each of these subjects was *levels-based* with the mark range divided into a number of 'levels' or 'bands'. Each band related to a particular range of marks and had an associated description of the criteria that were expected to be met at that level.

Semi-structured interviews were conducted with 11 GCSE teachers. The participants were one English/English Literature teacher, five Geography teachers and five ICT teachers. The interview questions asked participants to describe how internal moderation is conducted and the thought processes involved, including how marking guidance is used and whether they feel the process works well.

Four mock internal moderation sessions were observed with some of the same participants: one session involving GCSE English/English Literature (one teacher); one session involving GCSE ICT (two teachers); and two sessions involving GCSE Geography (three teachers and two teachers respectively). At the English teacher's school, internal moderation of coursework was usually carried out by the head of department so, in order to mimic this, two of his colleagues also conducted some marking and these marks (along with some of his own marking) were moderated by the head of department as an individual activity. For Geography and ICT, the internal moderation was carried out as a group activity, mimicking usual practice in these school departments.

The sessions used student coursework provided by the researcher with each teacher marking four different students' projects before the internal moderation session. The students who prepared the coursework were unknown to the teachers, representing a departure from the usual situation where teachers mark work from their own students. However, during internal moderation teachers usually evaluate work from some students that they teach and some who are taught by colleagues, so it is not unrealistic to ask teachers to mark work from students they do not teach. Where mock internal moderation was an individual exercise (English) the time available allowed all 12 coursework folders to be considered. In all other cases (i.e., all those where the internal moderation exercise was conducted as a group activity), six coursework projects were considered in each case. (Note that this is more a reflection that the English moderation session was carried out after the school day when the participant had more time available, than an indication that individual moderation is faster or more efficient.)

All sessions were observed by the researcher and audio-recorded. For the individual session with the GCSE English participant, he was asked to 'think aloud' whilst conducting the task. He was instructed as follows, based on Ericsson and Simon (1993): 'I would like you to say out

loud everything that you would normally think to yourself silently whilst you are moderating. It may help if you imagine that you are in the room by yourself.' There is some debate around whether the 'think aloud' method can affect a participant's thinking whilst conducting a task (e.g., slowing down normal processes), however, it is generally felt to be a useful method providing more information than observation alone (for further discussion see Crisp, 2008; Ericsson & Simon, 1993; Green, 1998; Kobrin & Young, 2003; Nisbett & Wilson, 1977).

The questionnaire data reported in this article comes from a longer questionnaire that addressed teacher marking as well as internal moderation (Crisp, 2013) and which was completed by 378 secondary school teachers from a range of subject areas across the Arts, Sciences, Humanities, Technology, English, Business and Social Sciences. Only teachers with experience of internal moderation were asked to complete the questions relevant to the current study, thus the numbers of respondents for the relevant questions were lower, ranging from 261 to 282 (with a total of 288 answering at least 1 of the questions relating to internal moderation). Of the 288 responding teachers, 158 taught GCSE (but not A level), 54 taught A level (but not GCSE) and 68 taught both[1]. The relevant questions covered use of marking criteria in internal moderation, differences between internal moderation and marking, and any effects of social interactions and group dynamics on the process.

It should be noted that there are some limitations to this research. Firstly, the number of teachers involved in the interviews and internal moderation was fairly small. This was necessary due to the in-depth nature of analysis needed, but it is possible that variations in practice might have been seen if different teachers had participated. Secondly, the 'mock' nature of the internal moderation sessions could be criticised on the grounds of not being as authentic as asking teachers to evaluate the work of their own students. Work from students unknown to the teachers was used to avoid any risk of the research affecting the 'live' marking and internal moderation process for the schools' own students. The use of work from students not known to the teachers could mean that some specific issues around assessing their own students are missing from the current data. However, as mentioned earlier, during internal moderation teachers usually look at work from some students that they have not taught, as well as some from students that they have taught, so it is hoped that using students unknown to the teachers is not a significant weakness to the method.

Findings

Insights from the interviews

During interviews, teachers were asked about the process of internal moderation at their school. The English teacher described the individual approach to internal moderation at his school and how he collects up all marked coursework for the subject and then checks a sample of each teacher's marking. All other teachers interviewed described their use of *group moderation* with each teacher evaluating some examples of student work from other classes (e.g., a high-, middle-, and low-scoring example from each class might be selected for consideration and marked by the other teachers). This marking might be conducted before or at an internal moderation meeting where marks would then be compared between teachers, discrepancies discussed, justifications given and agreements reached. The internal moderation could result in one or more individual teachers returning to their marking for all coursework projects and adjusting their marks to bring them into line with the marking standards being applied by their colleagues. Most teachers felt that internal moderation worked well, and several quoted as evidence of this that their marking standards are usually similar (with only small mark differences found if any) and that their marks had rarely been adjusted by the external moderation process.

Teachers were also asked about any differences in how they evaluate work and in the use of criteria in internal moderation compared to marking. Mostly, the evaluation process was thought to be very similar between these two contexts. Some participants commented that during internal moderation each individual coursework project was considered more quickly, particularly if the second marking was conducted in the internal moderation meeting rather than in advance of it. In terms of use of the marking criteria, this was generally felt to be similar but a few teachers suggested that they made less direct use of the detail of the marking criteria when evaluating during internal moderation, as a broader view is taken. Some teachers mentioned that, when marking as part of internal moderation, they tended to be slightly harsher on students that they did not know because they had not seen the work progressing, and that they tended to defend the marks they had given to their own students. Nonetheless, the internal moderation process was thought to address any possible biases towards or against known or unknown students through discussion and refinement of marking.

Insights from the mock moderation sessions

The teacher participant who conducted the *mock moderation* as an individual activity considered each coursework folder in turn and usually orientated himself to the topic when starting to read. This was often followed by noting the mark(s) originally given to the work and what this may suggest (e.g., "a band 3 essay, this will probably not be as good as the previous piece"). Reading during internal moderation appeared to involve some skimming with any annotations (ticks, comments, etc.) somewhat guiding this process. An absence of teacher annotations, or only brief annotation, was sometimes commented on by the moderating teacher. Agreement with teacher marks or annotations was sometimes noted. In addition, the participant sometimes noted that work had been over- or under-valued, which then led to adjusting marks.

In the group moderation meetings, the teachers compared and discussed the marks, considering each coursework project in turn. For each project, they began by each stating the total mark that they had given. If the total marks were close together then little discussion was required but the grade that was likely to be equivalent to that mark might be noted. For a project with slightly larger differences between the total marks given by different teachers, there was a much lengthier discussion. One tendency was for the teacher who was furthest from the others to immediately consider their own marking to have been too lenient or too harsh.

At one school, the internal moderation meeting began by comparing the teachers' rank orders of total marks for the coursework projects. Any significant differences in rank orders were noted. The discussions around this process involved each teacher stating the mark they gave, comparing the mark to those proposed by their colleagues, noting similarities and differences and possible adjustments to marks.

1. A small number taught another qualification (e.g., BTEC Entry level) either as the only qualification they taught or alongside GCSE and/or A level.

There was a significant comparative element to the discussions in this school in terms of the teachers comparing the quality of one coursework project to another, often in terms of specific marking criteria. After identifying those coursework projects where there were discrepancies in the marks given by different teachers, the projects in question were discussed in more detail.

Discussion during the mock internal moderation meetings involved going back to evidence within the coursework projects, using the marking criteria (or a marking cover sheet attached to each project which lists the marking criteria), summarising the contents and features and quality of the work. Evaluations were usually stated at a fairly broad level (e.g., evaluation of data representation) but sometimes at a more specific level (e.g., evaluation of map use). Criteria with which there were discrepancies for a particular student were identified which led to discussion of different perspectives on that particular aspect of the student's work. Discussions sometimes focused on whether a particular part of a student's work constituted evidence towards a particular criterion. One teacher would show the other(s) the evidence of a particular criterion that they had accepted, and then the teachers would reach mutual agreement on whether to accept this as sufficient evidence. The more extreme-marking teacher might question their reasons for their mark and/or describe why they gave that mark and the other teachers would describe their rationales for the marks they gave. There were also discussions about the requirements of the marking criteria to clarify and confirm interpretations of these. This process led to agreement on the appropriate mark using the marking criteria. Usually marks were adjusted away from the more extreme mark and towards the consensus. The grade likely to relate to the mark was sometimes noted once the mark had been agreed or during discussions.

In two of the three mock group moderation meetings in this research, observing teachers' interactions suggested that the more senior teacher present tended to lead the direction of the discussion and appeared to be less likely to adjust the marks they gave, although they were not unwilling to listen and reconsider their initial mark. It is plausible that a more senior teacher has the most experience with marking and that their judgements are likely to be closest to the national standards. In which case, it would be appropriate that they have a stronger influence on the discussions and decisions. However, if their understanding of the marking criteria and expected standards is no stronger than that of their colleagues, then their more influential position could have an unhelpful effect on decision-making.

Insights from the questionnaire responses

As described earlier, the questionnaire data reported here comes from a longer questionnaire that addressed coursework marking as well as internal moderation (Crisp, 2013). Those teachers without involvement in internal moderation were asked to skip the relevant questions. Some 25 to 31 per cent of teachers omitted the closed questions in this section. This suggests that these teachers work in departments where they are the only teacher (perhaps due to small school size or limited uptake of the subject) or where one teacher, perhaps the head of department, conducts the internal moderation alone.

The questionnaire included three closed response questions on internal moderation with an open response question following each to elicit further detail. The closed response questions and the frequency of different responses to these are shown in Table 1.

Table 1: Closed response questions and frequencies of response

During internal moderation procedures, do you use the mark scheme criteria in exactly the same way as when marking? (N=282, omitted by 25.4% of whole sample)

Yes	No
96.8%	3.2%

How often do social interactions or group dynamics between teachers affect internal moderation procedures? (N=281, omitted by 25.7% of whole sample)

Never	Occasionally	Sometimes	Usually	Always
47.3%	27.8%	18.5%	4.6%	1.8%

Are there any other ways in which internal moderation judgements differ from marking judgements? (N=261, omitted by 31.0% of whole sample)

Yes	No
22.2%	77.8%

Firstly, teachers were asked whether marking criteria are used in exactly the same way in internal moderation as during marking. The majority of those responding felt this was the case. Respondents were asked to give examples if they felt there were differences. The responses are listed in Table 2. Comments included that internal moderation involves considering work more holistically, ranking work into order, using the criteria to justify decisions to others, and that teachers' annotations are used as well as the marking criteria during moderation.

Table 2: Reported examples of ways in which marking criteria are used differently in internal moderation compared to marking

- *Use the detailed breakdown, then look at how it is marked after.*
- *Rank the grades. Look again in coursework if think too low/too high.*
- *Also look at comments and cross-referencing.*
- *We need to compare decisions and justify them so I refer to it much more.*
- *Because we moderate our interpretations of what answers mean.*
- *Generally look at work as a whole.*
- *Use expertise of other teachers involved in marking/moderation.*
- *Sometimes we refer to teaching resources for the staff to help further.*
- *When marking I mark by question. When moderating I also mark overall.*
- *One member of staff is an examiner for an awarding body so she sometimes has additional information which can clarify the mark scheme.*
- *Don't focus on them.*
- *Using marking criteria as guidance.*
- *Take an overview; look at the annotations of the teacher to check where marks have been awarded.*

Secondly, teachers were asked about social interactions or group dynamics and how frequently these affect internal moderation procedures. Most respondents reported that these were infrequent influences on moderation. However, nearly a quarter reported that social interactions or group dynamics at least 'sometimes' affected procedures. Teachers were asked to provide an example, if possible. Fifty-seven responded with at least one point and their comments were analysed by grouping similar responses together (see Table 3). Some responses were

Table 3: Reported examples of social influences on internal moderation

Response	*Frequency*
Positive discussion to reach agreement/happy medium over differences in views/marks	6
Good relationship with other staff/positive working relationships	2
Constant ongoing dialogue with other staff during coursework teaching	2
Hierarchy in department	2
Level of experience or familiarity with qualification (e.g., inexperienced teachers led by more experienced, experienced teachers' marks get agreed more quickly)	5
Differences in experiences (e.g., different subject experiences)	4
Personality (e.g., persuasive, wilful, argumentative, emotional)	5
Collaborative working issues – need for give and take in team working	1
Taking offence/taking criticism badly	5
Personality clashes/personal differences	2
Risk of unprofessional behaviour	1
Taking over from another teacher who has not taught the group well	1
Differences in perceptions of student performance/differences in marks/differences in ideas about standards	10
Interpretation of the criteria (e.g., helps to hear how someone else interpreted the criteria)	4
Differences in thoroughness	1
Occasional bias against a pupil can be removed in internal moderation	2
Can consider the nature of the student group (e.g., if less able)	1
Social interactions affect time and support available	1
Practical issues regarding time (e.g., time-consuming perhaps because of arguments or getting off track; organising a time to suit everyone)	6
Total	**61**

positive, implying that working together was a useful and supportive part of the process. For example, six teachers commented that positive discussion was used to reach agreement over differences in views. Frequent ongoing dialogue with other teachers during coursework-related teaching was also mentioned as a positive feature. Several neutral comments about other teachers were made. These included that the level of experience of staff, differences in experiences such as different subject experiences and the hierarchy in a department could affect the internal moderation process (e.g., teachers with less experience may be led by teachers with more). Five teachers mentioned aspects of personality, such as persuasiveness or argumentativeness, as influences on internal moderation. Other comments included differences in perceptions of student performance or marks, and differences in interpretations of criteria. However, it was unclear from these comments how teachers felt social interactions in relation to these differences influenced the process. Several negative influences were mentioned, including issues about colleagues taking offence at criticism, and personality clashes. Student-focused comments included that the nature of a student can be taken into account through discussion, and that occasional bias against an individual student can be resolved. Two practical considerations were also noted: that social relationships affect the amount of time and support available to the teacher in relation to their marking; and that the process can be time-consuming due to arguments or getting 'off-track' during meetings.

The third closed question asked teachers if there were any other ways in which internal moderation judgements differ from marking judgements. Over three quarters of those who responded reported that there were not, suggesting that many teachers consider judgements in internal moderation similar to those in marking. Those that felt there were differences were asked to give an example, to which 38 teachers gave a response. Comments included: that different interpretations of the marking criteria influence the internal moderation process; that a view from a teacher who is less familiar with the student can aid objectivity; that internal moderation decisions involve discussion; and that one teacher may see qualities in the work that another did not identify.

Discussion

This study provides insights into the internal moderation processes used in schools to standardise marks before submission for external moderation. The mock internal moderation sessions showed that, as well as behaviours relating to the consideration of individual coursework projects (and thus common with marking), a number of additional behaviours occur, including noting and/or agreeing with the mark given or comments made, discussion of where evidence in the work meets particular marking criteria, discussion of requirements, and adjustment of marks. Interview comments suggested that internal moderation is felt to address any biases towards or against known or unknown students. In the questionnaire responses, teachers generally reported that internal moderation uses marking criteria in the same way as marking, though student work may be considered more holistically in the former. This may imply that the criteria provide not only the basis for judgements about marks in internal moderation processes, but also a common terminology that can be used by teachers in internal moderation discussions.

Given the levels-based nature of the mark schemes used to assess most non-examination work contributing to GCSEs and A levels, it is logical to expect that teachers look for evidence in the student work relating to particular skills, attempt to identify the most appropriate level by looking at the criteria described for each level, and then judge the mark to be awarded from the range relating to the level and how well the work has met the criteria. The current data would generally seem to be consistent with this, though perhaps there is insufficient data to claim this conclusively.

Previous research on grading meetings has shown that group dynamics influence the judgements of examining teams (Murphy et al., 1995). In the current context, teachers tended to report that social interactions and group dynamics were an infrequent influence on internal moderation. Whilst group dynamics were not felt to be a strong influence, there is clearly a social dimension to internal moderation. This is perhaps exemplified in the tendency for a teacher whose initial mark for a project was furthest from the other teachers' marks to immediately express that they were likely to be the one whose marking was out of line with national standards. Whilst this could be a logical

assumption in the circumstances, there is potentially a social element to this with confessing their own (possible) error before they are criticised by others acting as a device for 'saving face'. It would also seem to be a positive sign about the working relationships of the teams involved that participants were comfortable admitting a potential error to their colleagues. This relates to the notion of 'team psychological safety', which is defined as a shared belief amongst team members that the team is a safe context for interpersonal risk-taking (Edmondson, 1999). Team psychological safety facilitates behaviours such as admission and discussion of errors, and seeking information and feedback, and is associated with team learning.

As Adie, Lloyd and Beutel (2013) point out, the aim of a moderation processes is to provide 'a way to develop a shared understanding of standards of achievement and the qualities that will denote evidence of these standards' (p.971). Elements of this can be seen in the findings of the current study. Work by van der Schaaf, Baartman and Prins (2012) on moderation in a university context in the Netherlands analysed the quality of argumentation when tutors evaluated student portfolios. They found judgements to be of low-quality with many articulations not relating to relevant evidence. In contrast, the current study in the context of GCSE suggests considerable focus during internal moderation on relevant evidence in student work with frequent discussion of the location of relevant evidence and whether this evidence is sufficient to meet a particular criterion. This, along with reference to the marking criteria, discussion of requirements and the meaning of the marking criteria, is consistent with Cook and Brown's (1999) notion of tacit knowledge being made explicit and helping to refine and create new ways of knowing (a notion previously applied to consensus moderation by Klenowski and Wyatt-Smith, 2013). The new 'ways of knowing' created by involvement in an internal moderation meeting (and perhaps also to some extent from feedback on internal moderation in cases where it is conducted individually by a senior member of a department) should inform the remaining and future marking of each individual teacher in terms of understanding marking standards and how aspects of student work provide evidence of elements of the marking criteria.

Previous research has suggested that internal moderation is a useful professional development experience for teachers (e.g., Harlen, 2005; Smaill, 2012; Klenowski & Wyatt-Smith, 2010). The omission rate on the internal moderation sections of the questionnaire suggests that around a quarter of teachers who mark non-examined work may not get this experience, presumably due to being the only teacher in the school for a particular subject, or because one teacher conducts the internal moderation alone. This is an interesting finding in itself as, either through circumstance or design, these teachers are missing out on a potentially useful professional development experience.

Some of the potential challenges to teacher assessment may be at least partly mitigated by internal moderation. Purported challenges include that written criteria are subject to interpretation (e.g., Sadler, 1998), that teachers use tacit knowledge as well as the written criteria, and that they may assume that their own tacit knowledge is the same as that of others (Klenowski & Wyatt-Smith, 2013). Discussion of the meaning of criteria and of examples of how this is evidenced in student work would seem likely to reduce these problems.

Another outstanding question is whether both individual and group internal moderation approaches are equally effective. Group moderation would seem to have the benefit of discussion, of jointly refining understandings, and greater potential for continuing professional development. However, if one experienced teacher has a good 'feel' for the standards expected and a good understanding of the criteria, it could be easier and/or more efficient to obtain a coherent rank order for all students taking a particular subject at a particular school through one teacher working alone to moderate the work. This might provide weaker development for the other teachers but, arguably, achieving accurate results for the current cohort of students is a more immediate aim of internal moderation than providing professional development that may aid future practice. Further research could usefully explore the relative success of group and individual approaches to internal moderation in terms of whether a school's marks provide an appropriate rank order, whether a school's marks are adjusted, and whether individual teachers become more aligned with national standards over time through their experiences or feedback from moderation.

The evidence gathered in this research does not suggest any significant problems with the nature of the internal moderation processes used in schools in relation to non-examined GCSE and A level work. Attention is paid to relevant evidence in student work, moderation is reported to be infrequently influenced by group dynamics, the process is thought to act to remove any potential personal bias, and teachers tend to report that the process works well.

References

Adie, L., Lloyd, M., & Beutel, D. (2013). Identifying discourses of moderation in higher education. *Assessment & Evaluation in Higher Education, 38*(8), 968–977.

Allal, L., & Mottier Lopez, L. (2014). Teachers' professional judgment in the context of collaborative assessment practice. In C. Wyatt-Smith, V. Klenowski & P. Colbert (Eds.), *Designing assessment for quality learning* (pp.151–165). Dordrecht, The Netherlands: Springer.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy & Practice, 18*(4), 451–469.

Cobb, P., Gravemeijer, K., Yackel, E., McClain, K., & Whitenack, J. (1997). Mathematizing and symbolizing: The emergence of chains of signification in one first-grade classroom. In D. Kirshner & J. A. Whitson (Eds.), *Situated cognition: Social, semiotic, and psychological perspectives* (pp.151–233). Mahwah, NJ: Lawrence Erlbaum.

Cook, S. D. N., & Brown, J. S. (1999). Bridging epistemologies: The generative dance between organizational knowledge and organizational knowing. *Organization Science, 10*(4), 381–400.

Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation, 13*(5), 401–434.

Crisp, V. (2008). The validity of using verbal protocol analysis to investigate the processes involved in examination marking. *Research in Education, 79*(1), 1–12.

Crisp, V. (2013). Criteria, comparison and past experiences: How do teachers make judgements when marking coursework? *Assessment in Education: Principles, Policy & Practice, 20*(1), 127–144.

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

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly, 44*(2), 350–383.

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

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge, UK: Cambridge University Press.

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education, 20*(3), 245–270.

Johnson, S. (2013). On the reliability of high-stakes teacher assessment. *Research Papers in Education, 28*(1), 91–105.

Klenowski, V., & Wyatt-Smith, C. M. (2010). Standards, teacher judgement and moderation in contexts of national curriculum and assessment reform. *Assessment Matters, 2*, 107–31.

Klenowski, V., & Wyatt-Smith, C. M. (2013). *Assessment for education: Standards, judgement and moderation*. London: Sage.

Kobrin, J. L., & Young, J. W. (2003). The cognitive equivalence of reading comprehension test items via computerized and paper-and-pencil administration. *Applied Measurement in Education, 16*(2), 115–140.

Murphy, R., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J., & Gower, R. (1995). *The dynamics of GCSE awarding: Report to the School Curriculum and Assessment Authority*. Nottingham: University of Nottingham.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–259.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice, 5*(1), 77–84.

Sadler, D. R. (2013). Assuring academic achievement standards: From moderation to calibration. *Assessment in Education: Principles, Policy & Practice, 20*(1), 5–19.

Smaill, E. (2013). Moderating New Zealand's National Standards: teacher learning and assessment outcomes. *Assessment in Education: Principles, Policy & Practice, 20*(3), 250–265.

van der Schaaf, M., Baartman, L., & Prins, F. (2012). Exploring the role of assessment criteria during teachers' collaborative judgement processes of students' portfolios. *Assessment and Evaluation in Higher Education, 37*(7), 847–860.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge, UK: Cambridge University Press.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. J. (2010). The centrality of teachers' judgement practice in assessment: A study of standards in moderation. *Assessment in Education: Principles, Policy & Practice, 17*(1), 59–75.

Which students benefit from retaking Mathematics and English GCSEs post-16?

Carmen Vidal Rodeiro Research Division

Introduction

Following the Wolf Report (Wolf, 2011), the UK Government legislated that from September 2013 all young people who did not achieve a grade C in Mathematics and English General Certificate of Secondary Education (GCSEs) had to continue studying these subjects post-16. Therefore, since 2014, students failing this requirement have continued to work towards achieving these qualifications or an approved interim qualification as a 'stepping stone' towards a GCSE. For some students, reaching the GCSE standard may potentially have required progressive stepping stones, for example, through Functional Skills qualifications, or through Foundation and Higher Free Standing Mathematics Qualifications.

According to a report published by the Policy Exchange in summer 2014 (Porter, 2015), 27% of the cohort who took GCSE English did not achieve a grade C or above (just over 125,000 students) and 31% of the cohort who took GCSE Mathematics did not achieve a grade C or above (just below 180,000 students). These students, who should have retaken English and Mathematics post-16, could also have been studying a variety of different courses. Some could have gone on to study academic courses, such as General Certificate of Education Advanced Subsidiary/Advanced levels (GCE AS/A levels), some could have been following alternative courses at different levels, such as BTECs, Cambridge Nationals, Cambridge Technicals, or vocationally related qualifications, and some might not have taken any other qualification.

Changes to the funding policy for 16–19 students in state-funded schools and colleges (for details, see https://www.gov.uk/guidance/16-to-19-funding-how-it-works) and the reform of post-16 accountability measures (DfE, 2017) are likely to have had an impact on enrolments in these centres and on entries for all types of qualifications in Key Stage 5 (KS5), but in particular for GCSEs in English and Mathematics. The 2015/16 academic year was the first in which it became a condition of colleges' funding that students who had previously achieved a grade D in English or Mathematics should retake the qualification. As a result, the overall number of entries among students aged 17 and over increased (Ofqual, 2016; 2017).

Recently, educational bodies across the sector, for example, The Office for Standards in Education, Children's Services and Skills (Ofsted), (Burke, 2016; Exley, 2016); the Association of Employment and Learning Providers (Martin, 2017); the Association of Colleges (Exley & Belgutay, 2017); the National Association of Head Teachers (NHAT, 2017); and the Learning and Work Institute (Belgutay, 2017) have been calling for a change in the resit policy. Their main reasons for requesting a review of the policy include:

- concerns over the lack of resources across the education system due to the increasing number of students required to retake the qualifications (e.g., insufficient funding; pressure on staff; logistical issues). This is a particular challenge for further education (FE) colleges, where the majority of the students retaking English and Mathematics GCSEs are enrolled;
- the huge numbers of learners aged 17 and older who failed to improve their grades after resitting GCSEs in English and/or Mathematics. In fact, the 2015/16 Ofsted Annual Report (Ofsted, 2016) stated that many students were still not getting at least a grade C by the age of 19;