Vol 11 | Issue 2 (2023) | pp. 30-46

JOURNAL OF
Perspectives in Applied Academic Practice

https://jpaap.ac.uk/

# Searching for best practice in the application of mark schemes: Top ten tips for successful post-hoc marking

**Simon Lock, Lucy Berthoud, and Becky Selwyn, University of Bristol**

ABSTRACT

Uncertainties within most teaching and learning contexts make it difficult for assessment designers to predict the challenge experienced by students during an assignment. For this reason, 'post-hoc' adaptation of mark schemes sometimes take place *after* an assessment has been completed. This is often to compensate for inaccurate assumptions embedded within the assessment itself. Such adjustments occur within many spheres of education and our goal here is to open conversations around the use of these strategies. The overall aim of this paper is to explore and investigate some of these practices and propose a set of guidelines for educators in designing flexible mark schemes and achieving adaptable, yet consistent, marking processes.

**Keywords:** Post-hoc marking, mark scheme adaptation, assessment guidelines

## Introduction

The aim of this paper is to explore the topic of 'post-hoc' mark adaptation, that is: the flexible interpretation of predefined mark schemes to manipulate a mark profile to adjust it to within 'acceptable' bounds. This process involves working within the constraints of predefined marking criteria, whilst at the same time achieving flexibility to allow mark profiles to be adjusted (either up or down). This flexibility is essential because it is often impossible to determine the relative challenge or difficulty of an assessment at the time it is created. Various uncertain or unknown factors within the teaching and learning context will impact students' ability to successfully achieve when completing an assignment. By their very nature, such emergent factors are not known when the assignment is created. Such uncertainties can include:

1. Quality of course materials: especially the impact of changes in those materials (such as a switch to pre-recorded video);
2. Effectiveness of content delivery: especially the impact of changes in lecture delivery (such as switch to online teaching);
3. Student engagement with materials: especially the impact of changes in consumption (such as move to remote learning);
4. Effectiveness of small group tuition: especially the impact of changes in tuition (such as switch to online teaching sessions);
5. Effectiveness of personal study: especially the impact of health, mental health and personal issues.

6. Suitability of assessments: especially the impact of change in assessment (such as move to "open book" exams);
7. Suitability of assessment platforms: especially the impact of changes in platform (such as automated marking tools);
8. Extent of plagiarism and collusion: especially the impact of changes in assessment (such as a switch to online exams); and
9. Intrinsic challenges: the impact of limitations of academic staff knowledge (such as understanding of the curriculum).

The impact of the Covid-19 pandemic illustrates the wide range of change and uncertainty in all the above factors. There are, however, many other phenomena which can also result in variation of these aspects of the teaching and learning context, for example: turn-over of teaching staff, the introduction of new topics and modules, the introduction of new tools and processes etc.

For all these reasons, mechanisms are sometimes needed to allow markers to adjust the mark profile *after* an assessment has taken place. Such strategies are in common usage in many areas of education, for example the application of scaling by national exam boards (Pearson, 2023) as well as the use of comparative judgement in setting grade boundaries (Benton et al. 2022). The UK government's Office of Qualifications and Examinations Regulation (Ofqual) often refers to such activities under the umbrella of "maintaining standards" (Newton, 2020). It is the aim of this paper to engage in a conversation around the use of these strategies and propose a framework to better understand and critically assess the operation of these processes.

The work described in this paper was originally initiated as a response to Covid 19. The additional stresses placed on assessment by the pandemic caused the authors to reflect upon the approaches traditionally used to achieve flexibility in the application of mark schemes, as well as new approaches being adopted. It is hoped that the outcomes of this work will be valid at any time – the observations made are applicable wherever uncertainty exists and marking flexibility is needed.

## Literature review

Assessment of student learning is an important part of education. At its most basic level, assessment evaluates whether a student has met the learning outcomes of a course. It has also been argued that under certain circumstances assessment can be used to support learning, to provide feedback to students (e.g. formative assessments), to ensure academic standards are maintained, to clarify expectations (e.g. through constructive alignment), and for public accountability (Biggs, 1996; Brown, 2005; Millar, 2013). While development of the assessment itself to satisfy all these requirements is a complex and important topic of study, this paper focuses on the use of mark schemes to turn assessment performance into a grade.

The Cambridge Dictionary defines a mark scheme as "a plan for giving marks for students' answers in an examination" (Dictionary, 2021). Mark schemes, like assessments, should be both valid and reliable. A valid mark scheme awards marks to responses that demonstrate attainment of relevant learning outcomes, while a reliable mark scheme ensures consistency in marks awarded such that that any individual assessor awards similar marks to similar responses, and multiple assessors award the same mark to any individual response (Pollitt et al., 2008; Wiliam, 2001).

Mark schemes generally fall into two common categories: criterion-referenced or norm-referenced. Norm-referenced schemes compare students against other students taking the same assessment to produce a ranking from best to worst. These types of mark schemes do not demonstrate absolute student achievement and are therefore not appropriate when the aim of an assessment is to assess competence rather than relative ability. Criterion-referenced mark schemes compare answers to an external standard or set of criteria, making it possible for all students to achieve any mark, rather than being limited by their performance compared to others (Lok et al., 2016; Turnbull, 1989).

Most mark schemes have moved towards criterion-referencing rather than norm-referencing (Pownall & Kennedy, 2019), and use a range of quality assurance measures to ensure accuracy (e.g. using groups of experts to set the standard, testing the mark scheme on exemplars, training assessors in use of the resulting mark scheme, repeating the testing and training to iteratively improve it before deploying it, and moderation of outcomes). In spite of this, there are often problems associated with the use of mark schemes (Baird et al., 2004). Many reports find poor reliability between assessors (Bloxham et al., 2016) due to multiple reasons including the type of question: reliability is highest for numerical answers and worsens as response length increases (Black, 2019; Tisi et al., 2013). A logical suggestion is then to restrict the type of question used. However, this can lead to lower validity as the learning outcomes are not able to be assessed as completely (Stobart, 2009).

A well-designed mark scheme is one of the most important factors in reliable and valid assessments, but is a difficult skill to master and there have been calls for formal training, particularly for inexperienced academic assessors (Norton et al., 2019). A taxonomy of mark schemes has been proposed, to help assessors design mark schemes which allow discrimination between different answers, including allowing marking of unexpected responses. The top level of the taxonomy would be a mark scheme that provides principles of discrimination for good and bad responses, whereas lower levels may only provide examples of different responses (Ahmed & Pollitt, 2011).

However, it seems that regardless of how well designed a mark scheme is, assessor judgement is still required to apply it. An expert must use their judgement to compare a given response to the mark scheme to extract the appropriate mark. Even when well-designed mark schemes are used, assessors still use their judgement and intuition to award grades, often referring to the mark scheme to justify their view or to compare responses against each other (Brooks, 2012; Crisp, 2013). Moderation is widely used to ensure consistent application of mark schemes both over time for individual assessors and between assessors, but moderation cannot undo the effects of a poorly designed mark scheme.

In summary, mark schemes should be valid as well as reliable, and a good mark scheme usually requires iterative testing and improvement of drafts before final deployment. Increasing reliability can reduce validity if more constrained question types are used. Assessor training and judgement are also important in applying a mark scheme correctly. The rapid shift to online assessments since the start of the Covid-19 pandemic in 2020 has created a two-pronged problem. Firstly, many of the new assessments rely at least partly on automatically applied mark schemes to very constrained question types (e.g. multiple choice or calculated numeric questions), meaning there are possible problems with validity. Secondly, many academics lack experience in writing questions and mark schemes for open-book online assessments, so it is more difficult to design and apply a mark scheme. These combine to produce a real risk that mark schemes are unfair and do not assess student learning in the intended way. This paper attempts to address this risk by investigating how academics interact with mark schemes, particularly related to online assessments, and then by identifying some key advice for designing assessments and mark schemes for

uncertain times. It looks at ways in which assessor judgement can be added to online assessments, and how to best to design an assessment and mark scheme for this purpose.

## Methodology

This section describes the methodology used during this research project, an overview of which is illustrated in Figure 1. The project began with a focus group activity involving teaching staff to generate ideas for subsequent stages of the process. These ideas were then rationalised in order to generate a coherent set of marking strategies. These strategies were then embedded within a questionnaire which was sent out to key stakeholders within, as well as outside, our institution. In order to gain a broader perspective, and to ensure that any insights and possible conclusions were relevant to other institutions, this same questionnaire was also presented to delegates at the Horizons in STEM Higher Education Conference (Lock et al., 2021). A total of 45 respondents from inside and outside our institution completed the questionnaire. The results of this survey (both feedback and opinion) were analysed and interpreted to derive guidance for current and future educators. Key elements of this whole process are described in the following subsections.

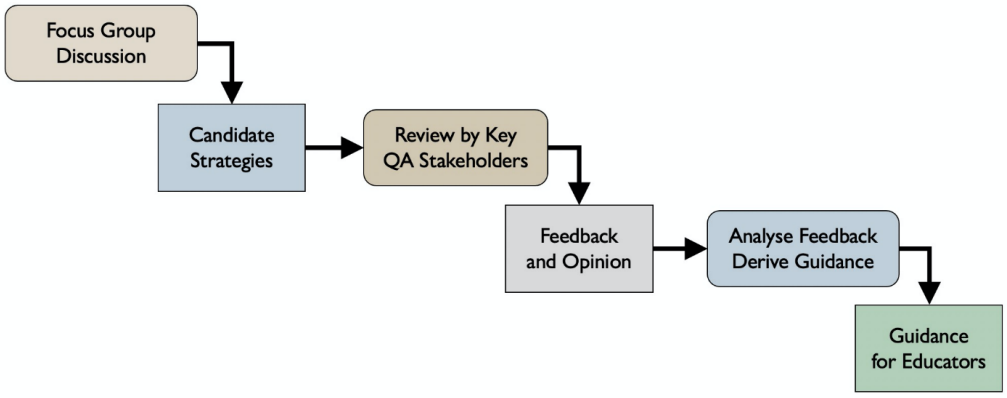


**Figure 1**. Methodology used in the current research (QA is Quality Assurance)

### Focus group

The aim of the focus group was to derive a set of currently used strategies and speculative practice in post-hoc marking. To this end, volunteer teaching staff members who fulfilled the following criteria were sought:

- Teaching for more than 2 years – some experience was desirable to give a context to the practice of marking.
- Teaching within the Faculty of Engineering – Engineering has a need for a particular mixture of numerical, theoretical and applied questioning.
- Experience of setting one or more online exams during the pandemic – some staff converted their exams to coursework and/or other types of assessment, so only those with specific experience of exam setting and marking were selected.
- Willingness to be open to explore their own and other marking practices – to generate different post-hoc marking practices, an open mind and awareness of possible bad practices was desirable.

Three members of staff were found for the focus group and discussion ran for a total of two hours. Notes were taken but the sessions were not recorded. The staff members comprised a professor and two senior lecturers from three different STEM disciplines with 35 years of teaching experience between them. The participants requested that they remain anonymous and that the delineation between current and speculative practice remain unspecified in the work. The ideas generated or 'candidate strategies' are outlined in more detail later in this paper. It is important to note that the focus in this research is on opening the debate and investigating different practices and their acceptability, and not on censuring staff members for their practice.

### Candidate strategies and questionnaire

The focus group identified many strategies including some which were purely speculative. They included strategies that the staff in the group had used in the past, those that they thought colleagues might have used and those that they thought unethical practitioners might use. A sift was then performed to filter out some of the more outlandish strategies. Similar or duplicate strategies were merged, and complex/composite strategies were decomposed into their elements, thus achieving a final set that were of similar scope and granularity. Finally, the strategies were divided into two groups: strategies to raise and strategies to lower student marks. The latter group were to a certain extent a mirror image of the first group, with a few exceptions.

A questionnaire entitled 'Appropriateness of Post Submission Moderation Strategies for Online Exams" with seven questions was designed. This aimed to explore the views of stakeholders on the strategies that were identified in the focus groups. The questions and answer options can be found in Annex A. As per good practice, options were balanced around a central point and an opportunity to express no opinion was offered. An ethics application was submitted to the Faculty of Engineering ethics committee and was approved (Application number 0241). The questionnaire was developed electronically and sent to a number of key stakeholders over a six-month period from March 2021 to October 2021. Note that this took place during the COVID-19 pandemic.

### Questionnaire respondents

To understand the stakeholders, it is useful to understand the exam setting and marking process at this institution. This process is a thorough and many-staged process which takes place over 4-5 months and is summarised in Figure 2. There are several more stages in the process not shown and the role of the central examinations office in organising, providing standards and checking has not been included for simplicity. Steps 1-3 are carried out by lecturing staff. The exams officer is a member of the lecturing staff of the department who is deputed to organise the exams. Step 5 is carried out by an external reviewer in the same subject from another University. The unit director is the member of lecturing staff who has taught the unit and set the exam (this can be more than one person of course).

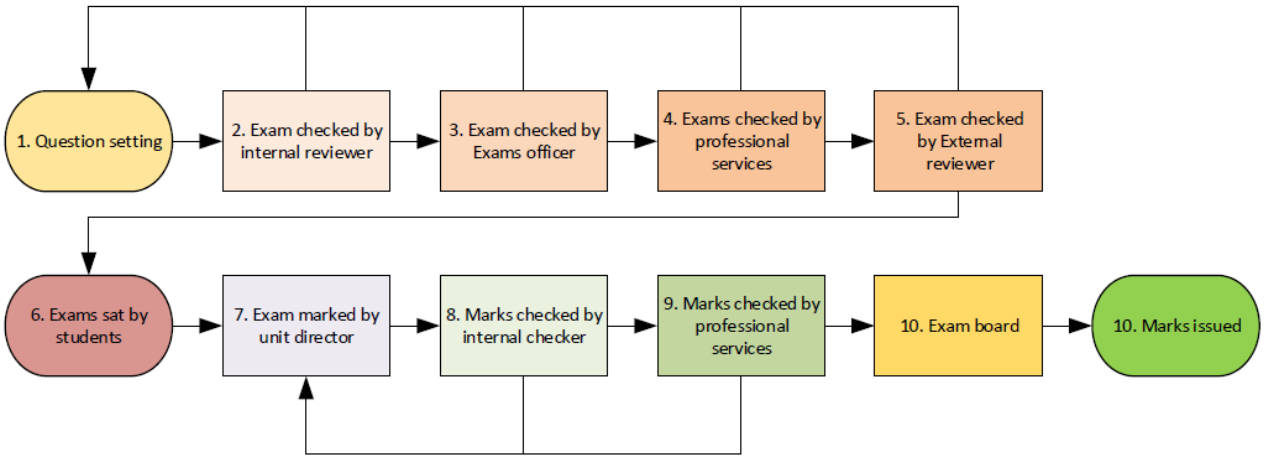


**Figure 2**. Exams setting and marking process at the author's institution (summarised)

In addition, the engineering degrees included in this exercise are accredited by UK wide industry-supported bodies called 'Institutions,' for example: the Institution of Mechanical Engineers. This accreditation process involves a close involvement of the Institution with the curriculum and assessment mechanisms, including the auditing of assessment processes, as part of a regular inspection visit to the University. For these reasons, it was considered useful to include accrediting staff as stakeholders. Therefore, the full range of stakeholders included in the study were:

- 33 Lecturing staff – staff who teach on the unit and who have set the exam.
- 1 Internal QA staff– staff in the University Quality assurance office.
- 3 Internal Exams officers – department member who organises exams and reviews mark profiles.
- 5 External examiners – staff at another University who check the exams for this University.
- 3 External accreditors – members of a UK institution who accredit the degree.

## Results and discussion

The questionnaires asked participants for their opinion of the set of strategies identified from the focus group activity. In addition, participants had the opportunity to provide open-ended feedback relating to the general concept of post-hoc marking. These different aspects of the questionnaire are presented in the following two subsections.

### Opinions of specific strategies

Figure 3 below illustrates the results of the survey of 45 participants. The question was "What is your view on the use of the following strategies to increase/decrease the marks of students".

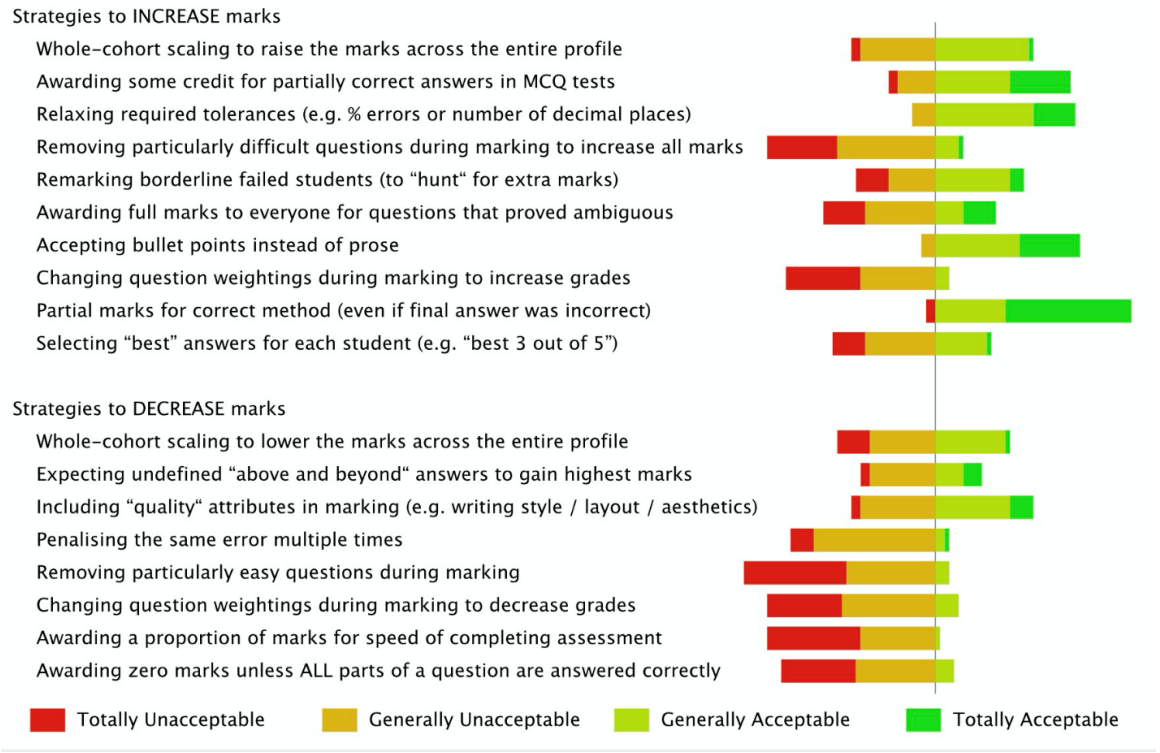


**Figure 3.** Graphical representation of perceived acceptability of moderation strategies

There is an intriguing diversity of opinion amongst the participants in our survey. Although there is universal agreement for some strategies, for others the perceived acceptability varies greatly. Almost all participants agreed that awarding partial marks for use of the correct method was acceptable. In many ways this is expected, given the common and widespread use of this practice. In contrast to this, the awarding of some credit for partially correct (i.e., 'grey area') answers in multiple choice assessments received acceptability ratings across the entire range from "Totally Acceptable" through to "Totally Unacceptable". It is interesting to reflect that for this strategy, respondents' perspectives varied depending on their role in the assessment process: External Examiners viewed this strategy as unacceptable, whereas Internal Exams Officers typically viewed it as acceptable. This may reflect the attitudes prevalent in different institutions or alternatively may be indicative of the rigour with which different assessment roles are undertaken. It is worth noting that discipline and educational background may also have an influence on respondents' attitudes and expectations.

Except for a few well-established and accepted common practices, most strategies typically had at least one participant who found it unacceptable (either "generally" or "totally"). Some strategies (for example the removing of particularly easy questions from the marking process) were generally agreed to be unacceptable. However, a small minority of respondents were at odds with the consensus. For these kinds of strategies, typically one or two participants disagreed with the majority. It is important to highlight however that this was not always the same person.

## Open-ended feedback

A reflexive thematic analysis was carried out on the answers to open-ended survey questions (Braun & Clarke, 2020). During an iterative process, data was coded, then codes were drawn together to form themes. This process was repeated several times until the data was satisfactorily coded and final themes were agreed. Four main themes were developed from the coded data – these are described below in the following four subsections.

### *Fairness for students*

Many respondents identified concerns around perceived fairness to students, noting that any change to the assessment could be deemed to be unfair if students were not aware of it pre-assessment. Respondents admitted that many of the strategies could be used as part of a predefined mark scheme, as long as students were made aware in advance, e.g. "*Changing the "rules of the game" when marking does not feel right or fair to the students"* and "*Many of these can be built into assessment design in an ethical manner, but would be totally unethical to implement in an ad-hoc manner".* Having clear, fixed assessment criteria that students can understand is a necessity for improving student assessment literacy and achievement (e.g. Howell, 2013; Smith et al., 2013), something which is not achievable if the goalposts are moved post-submission.

Fairness for students was also evident in that strategies to increase marks were far more acceptable to the respondents than strategies to decrease marks. This generosity towards students was reinforced by text comments: "*If the marks are too high, it means the questions were too easy, it's not the fault of the students and they should not be penalised"* and "*I expect that we often find many ways to justify increasing marks, as we seem to have a strong aversion to failing students"*.

There is a pull factor in that students themselves would no doubt be happier with higher marks, but higher education is also different to schools and further education in that their staff teach, set the exams, and mark them. This may also produce a bias in wanting their students to do well. To counter this, staff are continually urged to make use of the full range of marks and not to be afraid of reducing marks (where appropriate) to achieve more accurate mark profiles and avoid the potential danger of grade inflation.

Another complication arises because grade boundaries within many disciplines in higher education (such as engineering and sciences) are often perceived as being fixed – there is an implicit minimum 'skill level' standard that students are expected to reach. Within disciplines such as Medicine and Dentistry however, it is accepted that the varying difficulty of set assessments requires the appropriate calibration of mark boundaries (GMC, 2011). Within disciplines in which grade boundaries are fixed, the lack of flexibility means that if an assessment has been unexpectedly harder for students, the only way to compensate is to adapt the mark scheme post-hoc. This links to the positive perception of so many of the proposed strategies for post-hoc adaptation: it is the only way that many academics feel they can compensate for unexpected outcomes in assessments.

### *Expected mark profiles*

While only one comment quantified an 'expected' profile ("*The quickest way to make adjustments is to provide a piecewise correction of the whole cohort to take the marks within acceptable bounds (for example mean between 57.5% and 72.5%)"),* many referred to expectations implicitly. For example, "*If module marks turn out to be considerably out of the normal range, then typically a scaling policy is applied"* and *"there needs to be clear evidence of a defect (e.g.: question wrong, marks clearly adrift of a published standards)*". Further comments were even less specific, highlighting that post-hoc adaptations of mark schemes could be

used "*If the question given is too difficult and nobody gets it right."* Of course, in the ideal world, post-hoc adaptation would not be required because "*with a rigorous internal moderation process used to review module assessments, then the problematic questions, (particularly easy or difficult) should be weeded out or modified appropriately prior to the examination being issued."* The notion of an 'expected' mark profile has also been identified in the literature, showing that markers were most affected by a desire to keep mark distributions within the 'normal' range, while acknowledging the subjectivity of the acceptable range (Pownall & Kennedy, 2019; Yorke et al., 2000).

However, participants also acknowledged that there are a range of factors affecting mark distributions, not all of which can be accounted for during pre-assessment moderation: "*If the final marks do not fit the average criteria, there will be multiple reasons, some of which depend on circumstances that are outside the control of the examiners",* and "*The range of problems is difficult to envisage".*

This leads to the third theme – the processes surrounding post-hoc adaptations of mark schemes.

#### *Processes of post-hoc adaptations*

Participants highlighted the importance of any post-hoc adaptations following clear and transparent processes: "*circumstances must be taken into account, rather than arbitrary moving boundaries, being lenient or less so on specific questions. This is a matter for discussion with the programme director*" and "*Whole cohort scaling 'up' or 'down' is sometimes considered at module assessment board (MAB) meetings (although within strictly confined limits) for a particularly easy or difficult examination overall*." These comments are consistent with standard university regulations around moderation of assessments, which define how post-assessment, pre-exam board moderation should take place and be reported to exam boards (e.g. Bristol, (2021)). However, one participant also pointed out that: "*The quality of the cohort should not cause moderation, only the quality of the assessment"* and this raises the question of exactly how poor-quality assessments or unexpected mark profiles can be identified, and who is responsible for agreeing them. In cases where "*the university made a mistake (e.g.: impossible question, mistake in exam paper) then pretty much anything in students' favour is acceptable*," but many cases are less clear-cut than an error in the question and relate more to subjective 'expected' marks.

#### *Contingent factors*

The fourth and final theme developed by the analysis was that surrounding contingent factors make strategies seem either more or less acceptable. In some cases, participants did not feel able to decide whether a strategy was acceptable or not because *"some of the scenario[s] need more context to establish if they could be acceptable in some circumstances*." There was a particular interest in knowing the size of the cohort to inform decisions: *"There is a difference between big cohorts and small cohorts"* and *"I said whole cohort scaling was acceptable, but I would caveat that with it having to be a reasonable sized cohort (say 100+ students)"*. Another factor was the intention of the post-hoc adaptation: *"This all depends on intention. If the increase is due to errors in the assessment, ambiguity, poor QA, or misplaced difficulty, then it's more justified. In the neutral context of just 'increasing marks' it isn't*."

Ideally there would be no need for any post-hoc mark scheme adaptation to take place. However, as observed earlier and as seen in the results of the survey, this is not a perfect world. Various uncertainties in the teaching and learning context mean that markers inevitably must find ways of incorporating flexibility into their marking process. As one of the respondents in our survey put it: "*a posteriori mitigation by unit directors should be avoided*," and another says: "*changing the "rules of the game" when marking does not feel right or fair to the students.*" However, the use of adaptation strategies is commonplace and sometimes unavoidable.

While rigorous internal moderation to avoid problems in assessments is a good ideal to strive for, it assumes that internal and external checkers have perfect knowledge of the subject area, spot every mistake and that nothing changes between the moderation process and the assessment activity. While we agree with the suggestion that internal moderation is vital, we would also argue that deficiencies that exist in other aspects of the teaching delivery process (content delivery, tuition, software support tools etc.) should also be addressed. We do, however, fully support the principle of never allowing variations in the ability of the cohort to be the motivating factor for mark adaptation.

On a wider scale, we must also endeavour to maintain consistency and fairness between assessments within a year and between cohorts across academic years. During the focus group sessions, the complex notion of consistency was deconstructed by the participants into several separate aspects:

- Historical consistency: ensuring that standards of achievement are maintained with previous cohorts of students (to ensure that assessment gets neither 'easier' nor 'harder' over time);
- Programme consistency: ensuring that standards and expectations are similar across all units/modules in a programme (to ensure that outlying 'easy' or 'hard' units do not exist);
- Disciplinary consistency: ensuring that assessment standards are maintained relative to the expectations of the discipline (to ensure that graduates meet the threshold requirements of the profession);
- Institutional consistency: ensuring that standards of achievement are maintained across different programmes in the institution (to ensure that the institution maintains its reputation for quality).

These different influences provide opportunities to calibrate post-hoc marking. Specifically, they provide targets and benchmarks with which to guide the application of marking strategies to ensure fair and consistent outcomes.

## Top ten tips

To make our findings more easily accessible to other educators, we have distilled all that we have learnt from our focus groups and questionnaire into a set of 'Top Ten Tips'. The aim of this is to provide practical advice to guide those attempting to design flexible assessments and apply strategies for post-hoc marking. We emphasise that University guidelines should be followed, and these tips should only be implemented if they are allowable within the organisation's examination procedures.

Some of these tips were identified as a consequence of reflecting on the numerical data from the survey, others were created as a response to comments from the survey participants. Although based on a limited data set, it was still possible to identify situations where views converged and conversely where they diverged. Several of the tips we present are even more fundamental and advocate a mindset and attitude towards marking that we see as essential to enable flexible and consistent post-hoc marking to take place.

Our Top Ten Tips are as follows:

(1) **Embrace uncertainty**: Perfect prior knowledge is an elusive ideal that is rarely available to educators. It is important that we acknowledge the existence of uncertainty and accept the need

for post-hoc mark adjustment. Once we accept that it exists and occurs on a regular basis, we can put mechanisms in place to ensure consistent and effective use of adaptive marking mechanisms.

(2) **Everything is relative**: The changing educational context within which assessment takes place is a bed of shifting sands. For this reason, we must be flexible and responsive - recognising and rewarding *relative* achievement, rather than continuing to apply historic scales of *absolute* attainment (which may no longer be appropriate or valid).

(3) **Don't expect universal agreement**: No matter which strategy you select for post-hoc mark adaptation, you are bound to encounter someone who will not approve. The best you can hope for is 'local' acceptability - where the approaches selected are consistent with institutional attitudes and policies within which they are applied and operate.

(4) **Design for adaptation**: Plan ahead and try to design assessments that are 'susceptible' to post-hoc marking. If a set of acceptable adaptation strategies are known in advance, then the style, structure and content of the assignments can be crafted to accommodate such strategies. As a marker you may welcome the flexibility that this forethought provides - especially if uncertainties are present in the learning context and these have an impact on the planned assessment.

(5) **Ensure suitable challenge**: As we have seen from the results of the survey, strategies aimed at increasing marks are seen as more acceptable than those for reducing them. The implication of this is that it is wiser to lean towards more difficult and challenging assignments (with the potential to raise marks post-hoc) rather than opting for easier and more simplistic assessments (the marks for which may have to be subsequently lowered)**.**

(6) **Most things have been tried before**: Experienced academics have been grappling with the issues and challenges of post-hoc marking for many years. It is likely that most strategies have been devised and used before. There may already be a wealth of knowledge and experience in your institution ready to be exploited.

(7) **Strive for Equality**: Adaptations should be careful not to favour one student over another - what you do to one student's mark, you should do for all students. This does not mean that all marks are increased or decreased by the same amount, but rather, the same criteria should be applied in deciding whether to make an increase or decrease. This issue is especially important if there is a team of markers all working in parallel to ensure consistent outcomes across the cohort.

(8) **Strive for Consistency**: It is essential to calibrate grades and boundaries to ensure that the marks produced by post-hoc adaptations are consistent with other assessments. As noted previously, this might include efforts to ensure consistency at the institutional level, within the discipline, at the programme level or historically. This does not always mean achieving identical mark profiles, but rather, ensuring that equivalent levels of achievement are rewarded consistently.

(9) **Ensure transparency**: It is essential that the anticipated use of post-hoc adaptation mechanisms is made known to staff and students alike and a clear record kept of any adaptations that were made to marks during the marking process. This should include not only the changes made (increases and decreases) but also the criteria used to make decisions and justifications for those criteria. This paper trail of adaptations and calibrations is essential in order to allow the final marks to be justified to all stakeholders.

(10) **Fairness is everything**: Remember that the main characteristic of assessment that underpins our investigation is the notion of 'fairness'. This concept incorporates many of the points highlighted

above and includes issues such as: recognition of achievement, acceptability of adaptation strategies, equality, consistency, transparency and so on. We must always keep the overarching notion of fairness in mind at all stages of the marking process. We should also remain aware of the fact that these various aspects are often contradictory and can conspire to pull us in different directions. Well, we never said it would be easy.

## Further work

This work has highlighted the practical approaches taken to marking by a range of academics involved in assessment and quality assurance processes in higher education. It is based on survey responses from 45 participants from a small number of UK institutions. It would therefore be useful to extend the research to other UK and international institutions to see whether the reported results hold true across a wider range of respondents. This work has taken place against a background of a rapid shift to online assessments (which were often open book). However, many of the strategies proposed in the survey and the 'top tips' developed are likely to be relevant to other types of assessment (e.g.: closed book, in-person assessments). As such, the work could be extended to cover different types of assessment to investigate whether similar conclusions are found.

There is also much debate around whether there should be *any* modification of marks post-submission at all. Some may argue that an unexpected mark profile is indicative of a failure in the assessment mark scheme design and application (i.e., the assessment and/or mark scheme were not valid). Others may argue that if the assessment is competency-based and has passed the usual quality assurance processes (typically both internal and external checks by qualified academics) any resulting mark profile should stand. Little literature relating to modification of mark profiles has been found, apart from University Codes of Practice and it would be interesting to find out how common this type of mark modification is within the sector. Student views were not sought as part of this project, so future work could also examine student perceptions of post-hoc mark scheme adaptation.

The top tips presented should also be evaluated in use, with an aim of finding examples of how they can be applied to different types of assessment, and what the impact is on assessment outcomes.

## Conclusions

This work has considered the topic of post-hoc mark adaptation: the reinterpretation of predefined mark schemes after an assessment has taken place. There was particular need for such flexibility during the 2020-21 academic year, due to the pandemic-driven switch to online open-book exams. However, our findings and the outcomes of this work are generally applicable, wherever uncertainty exists within teaching, learning and assessment processes. As part of this work, focus groups were held to identify a range of strategies that can be used to change marks during the marking process after exams have been taken. These strategies were listed in a questionnaire and sent to key stakeholders (including lecturing staff, exams officers, external reviewers, and accrediting staff). This activity was undertaken within the authors' own institution, as well as other UK institutions (using an education conference to present the work and gain insights from educators across the HE sector). The acceptability of these strategies was then rated by 45 stakeholders and their opinions on these mechanisms collected and analysed. Strategies for increasing marks were found to be more acceptable to respondents than strategies for decreasing marks. All stakeholders thought that awarding partial marks for use of the correct method was acceptable and

changing question weightings during the marking process was universally considered unacceptable. There was much variability in the responses from survey participants, with the use of adaptation strategies being viewed diversely by different stakeholders. A detailed consideration of the outcomes of our study has led to the development of a set of 'ten top tips', provided to promote best practice and support educators in applying the principles and practices of post-hoc mark scheme adaptation.

## Biographies

*Simon Lock* is a Teaching Fellow and Senior Lecturer in Computer Science at the University of Bristol. His past work has focused primarily on supporting complex socio-technical systems through the use of software support tools. His current area of research interest is in automated marking and learning analytics.

*Lucy Berthoud* is a Professor of Space systems engineering (Teaching Pathway) at the University of Bristol. She has led curricula reviews, been an external examiner and researched assessment, resilient curricula, and the use of digital tools in education.

*Becky Selwyn* is a teaching-focused Senior Lecturer in Mechanical Engineering at the University of Bristol. She works on projects that aim to support students to succeed through developing better teaching and learning practices, and creating a safe, supportive environment for students to make mistakes in.

## References

Ahmed, A., & Pollitt. A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice, 18*(3), 259-278. https://doi.org/10.1080/0969594X.2010.546775.

Baird, J-A., Greatorex,J., & Bell, J.F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy & Practice, 11*(3), 331-348. https://doi.org/10.1080/0969594042000304627.

Benton, T., Gill, T., Hughes, S., & Leech. T. (2022). A summary of OCR's pilots of the use of comparative judgement in setting grade boundaries. *Research Matters*, *33*, 10-30. https://files.eric.ed.gov/fulltext/EJ1343617.pdf

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347-364. https://doi.org/10.1007/BF00138871.

Black, B. (2019, August 12). *11 things we know about marking and 2 things we don't …yet.* The Ofqual blog. https://ofqual.blog.gov.uk/2019/03/05/14572/.

Bloxham, S., den-Outer, B., Hudson, J., & Price. M. (2016). Let's stop the pretence of consistent marking: Exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education, 41*(3), 466-481. https://doi.org/10.1080/02602938.2015.1024607.

Braun, V., & Clarke. V. (2020). Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and Psychotherapy Research, 21*(1), 37-47. https://doi.org/10.1002/capr.12360.

Bristol, University of. 2021. *Regulations and code of practice for taught programmes.* University of Bristol. http://www.bristol.ac.uk/academic-quality/assessment/codeonline.html

Brooks, V. (2012). Marking as judgment. *Research Papers in Education, 27*(1), 63-80. https://doi.org/10.1080/02671520903331008.

Brown, S. (2005). Assessment for learning. *Learning and Teaching in Higher Education, 1*, 81-89. http://eprints.glos.ac.uk/3607/1/.

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. https://doi.org/10.1080/0969594X.2012.741059.

Dictionary, Cambridge. (2021). *Mark scheme*. Cambridge University Press. https://dictionary.cambridge.org/dictionary/english/mark-scheme.

General Medical Council. (2011). *Assessment in undergraduate medical education.*

https://www.gmc-uk.org/-/media/documents/assessment-in-undergraduate-medical-education---guidance-0815_pdf-56439668.pdf

Howell, R. J. (2013). Grading rubrics: hoopla or help? *Innovations in Education and Teaching International, 51*(4), 400-410. https://doi.org/10.1080/14703297.2013.785252.

Lock, S., Berthoud, L., & Selwyn. B. (2021). Searching for best practice in the successful use of mark schemes: Top ten tips for successful post-hoc marking. *Horizons in STEM Higher Education Conference*, Online, 29th - 20th June 2021.

Lok, B., McNaught, C., & Young. K. (2016). Criterion-referenced and norm-referenced assessments: compatibility and complementarity. *Assessment & Evaluation in Higher Education, 41*(3), 450-465. https://doi.org/10.1080/02602938.2015.1022136.

Millar, R. (2013). Improving science education: Why assessment matters. In D. Corrigan, R. Gunstone & A. Jones (Eds.), *Valuing assessment in science education: Pedagogy, curriculum, policy* (pp. 55-68). Springer Netherlands. https://doi.org/10.1007/978-94-007-6668-6

Newton, P. E. (2020). *Maintaining Standards: During normal times and when qualifications are reformed*. Ofqual. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/936340/Maintaining_Standards.pdf

Norton, L., Floyd, S., & Norton. B (2019). Lecturers' views of assessment design, marking and feedback in higher education: a case for professionalisation? *Assessment & Evaluation in Higher Education, 44*(8), 1209-1221. https://doi.org/10.1080/02602938.2019.1592110.

Pearson UK. (2023). *Results: GCSE and A-Level scaling*. https://support.pearson.com/uk/s/article/Results-GCSE-And-A-Level-Scaling

Pollitt, A., Ahmed, A., Baird, J-A., Tognolini, J., & Davidson. M. (2008). *Improving the quality of GCSE assessment.* Qualifications and Curriculum Authority (QCA).
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/605192/0108_AlastairPollitt_et_al_Improving_the_Quality_of_GCSE_Assessment_final.pdf.

Pownall, I., & Kennedy. V. (2019). Cognitive influences shaping grade decision-making. *Quality Assurance in Education, 27*(2), 166-178. https://doi.org/10.1108/qae-04-2018-0040.

Smith, C.D., Worsfold, K., Davies, L., Fisher, R., & McPhail. R. (2013). Assessment literacy and student learning: The case for explicitly developing students 'assessment literacy'. *Assessment & Evaluation in Higher Education, 38*(1), 44-60. https://doi.org/10.1080/02602938.2011.598636.

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research, 51*(2), 161-179. https://doi.org/10.1080/00131880902891305.

Tisi, J., Whitehouse, G., Maughan, S., & Burdett. N. (2013). *A review of literature on marking reliability research (report for Ofqual).* https://www.nfer.ac.uk/media/2002/mark01.pdf.

Turnbull, J. M. (1989). What Is… normative versus criterion-referenced assessment. *Medical Teacher, 11*(2), 145-150. https://doi.org/10.3109/01421598909146317.

Wiliam, D. (2001). Reliability, validity, and all that jazz. *Education, 3-13* 29 (3): 17-21. https://doi.org/10.1080/03004270185200311.

Yorke, M., Bridges P., & Woolf H. (2000). Mark distributions and marking practices in UK higher education: Some challenging issues. *Active Learning in Higher Education, 1*(1), 7-27. https://doi.org/10.1177/1469787400001001002

**Annex A: Questionnaire Questions**

1. Q: I consent to my anonymised answers being used as data in a research project run by the University of Bristol. This data may be published as part of a research paper.
   A: Yes/No

2. Q: Select which of the following best defines your role.
   A: Lecturing staff, Internal QA, Internal Exams officer, External QA, External examiner, External accreditor

3. Q: What is your view on the use of the following strategies to raise the marks of students?
   Whole-cohort scaling to raise the marks across the entire profile
   Awarding some credit for partially correct answers in MCQ tests
   Relaxing required tolerances (e.g.: % errors, number of decimal places)
   Removing particularly difficult questions during marking to increase all marks
   Remarking borderline failed students (to "hunt" for extra marks)
   Awarding full marks to everyone for questions that proved ambiguous
   Accepting bullet points instead of prose
   Changing question weightings during marking to increase grades
   Partial marks for correct method (even if final answer was incorrect)
   What is your view on the use of the following strategies to RAISE the marks of students
   A: Totally unacceptable/Generally unacceptable/Undecided/Generally acceptable/Totally acceptable/No opinion

4. What is your opinion on the use of the following strategies to LOWER the marks of students
   Whole cohort scaling to lower the marks across the entire profile
   Expecting undefined "above and beyond" answers to gain highest marks
   Including "quality" attributes in marking (e.g., writing style, layout, aesthetics)
   Penalising the same error multiple times
   Removing particularly easy questions during marking
   Changing question weightings during marking to increase grades
   Awarding a proportion of marks for speed of completing assessment
   Awarding zero marks unless ALL parts of a question are answered correctly
   A: Totally unacceptable/Generally unacceptable/Undecided/Generally acceptable/Totally acceptable/No opinion

5. Do you have any other comments regarding the above?
   A: (Free text)

**Annex B: Questionnaire freeform question responses**

1. "As much as possible, a posteriori mitigation by unit directors, either way, should be avoided. Different staff members have different notions of what is ""difficult"" or ""worth"" a certain grade. In my own experience most staff tend to be very generous in favour of the students when interpreting the 21-point scale (or maybe I'm just too harsh :) ). Either way, we should provide more formal guidelines and provide training. I expect that we often find many ways to justify increasing marks, as we seem to have a strong aversion to failing students; That should change. In addition, the current marking scheme (UK wide) is extremely narrow, not allowing us to easily separate students' performance. This narrow marking range makes the overall marking more sensitive to errors, and therefore, increases the need to use mitigation if something goes wrong, e.g. a single 10-point question can make the difference between a 2.1 and a 1st. We could, also, only use very small number of marks per question, allow partial marking, and other marking criteria (as for online test), but they must be defined a priori of the exam. Although the role of exam officer lies within academics, I can imagine a non-academic with a role focus on exams would help make the overall process fairer and more uniform across the school; There is room for improvement and a lot of work. Exam Officer academics do not have enough free time to action these changes."

2. “Questions for online exams need to be set at the appropriate level, considering the specific circumstances of the exam. For example, will the online exam be open book or closed book, or will students be able to backtrack on a MCQ test to see all questions prior to having to answer them, etc. If module marks turn out to be considerably out of the normal range, then typically a scaling policy is applied. Normally, submission moderation should only be applied when the assessment itself has deficiencies. The quality of the cohort should not cause moderation, only the quality of the assessment. There is no totally acceptable answer, but whole scaling the cohort is probably the fairest (different scaling methods can be used). Changing the "rules of the game" when marking does not feel right or fair to the students.”

3. "I am very surprised at the options that have been proposed, and in most cases I find them unacceptable. They all go around to find ways in order to adjust the exam marks, and are all arbitrary, without any academic justification. The quickest way to make adjustments is to provide a piecewise correction of the whole cohort to take the marks within acceptable bounds (for example mean between 57.5% and 72.5%). If you correct upward uniformly, a 100% score cannot be rescaled. There are issues with the individual marking schemes. Changing this during the marking process is like changing the rules of the game after the game has started - this is not acceptable. The examination papers undergo a rigorous process of preparation, proof-reading, revision and approvals. It the final marks do not fit the average criteria, there will be multiple reasons, some of which depend on circumstances that are outside the control of the examiners. These circumstances must be taken into account, rather than arbitrary moving boundaries, being lenient or less so on specific questions. This is a matter for discussion with the programme director."

4. "Some of the options depend on context. For example, ""best 2 of 3"" is perfectly fine if you tell students at the start that's how the exam will go, but doing it after marks are in I'd only accept if there was a problem/mistake on the exam. Similarly, accepting bullet-point answers depends on whether students were told to write complete sentences resp. write an essay. My general rule is if the university made a mistake (e.g.: impossible question, mistake in exam paper) then pretty much anything in students' favour is acceptable. "

5. "Where we have (only very occasionally) applied scaling, it has tended to be by adding a multiple of mark*(100-mark). For example, adding mark*(100-mark)/240 is the simplest formula that leaves 0 and 100 fixed while raising 40 to 50 (as that is not possible with a linear formula and this is the unique quadratic to do so)."

6. “Very interesting questions to consider. I hope that External Examiners have been included in your survey. I think that the length of time (number of years) that an examiner has been setting and marking exams has a significant impact/effect upon outcomes. It may well be that examiner's expectations regarding outcomes are affected significantly with the passing of time and you might consider this as a part of your investigation. Whole cohort scaling 'up' or 'down' is sometimes considered at module assessment board (MAB) meetings (although within strictly confined limits) for a particularly easy or difficult examination overall. However, with a rigorous internal moderation process used to review module assessments, then the problematic questions, (particularly easy or difficult) should be weeded out or modified appropriately prior to the examination being issued.”