# Step Marking Consultation

## Summary


1. At recent meetings of the TEI Forum and the Common Awards Management Board, there has been discussion of whether Common Awards could move over to a 'step marking' scheme. We undertook to investigate the implications of a move before consulting the Common Awards community more widely.

2. In a step marking scheme markers are required to use only a subset of the percentage marks currently available to them. For instance, they might in the 60s be restricted to using only the marks 61, 63, 66 and 68.

3. This document explains, positively, that such a scheme has the potential to be more transparent and more efficient than our current marking practice, that it involves no significant loss of precision or reliability, and that it need not lead to significant grade inflation or deflation. It points to other universities, comparable to Durham, in which such a scheme is employed, and suggests that no change to Durham's systems or core regulations is required for implementation.

4. This document also explains, negatively, that the gains in efficiency involved might be more apparent than real, and that such schemes can generate amongst both markers and students the impression that marking has become less precise (whether or not that impression is justified).

5. The document finishes by explaining the current consultation, and posing some questions on which we are seeking feedback from staff and students.


## The case for step marking

6. Our marking criteria (and the marking habits that they reflect and reinforce) provide a way of translating qualitative judgments into numerical marks – but this translation is an entirely conventional affair, and in one sense entirely arbitrary. We choose, for instance, to set the boundary between upper second and first class the at number 70. We could have set it at the number 3, the number 270, or any other number. The number 70 has no direct meaning: it does not mean that

70% of the learning outcomes were met, not that first class work is at least $^{70}/_{40}$ times as good as a bare pass.

7. There is no reason to suppose that markers are able reliably to distinguish one hundred different bands into which work might fall – or fifty (if we mark, in practice, between about 35 and about 85). We should be wary of letting the numbers we use encourage a spurious assumption of accuracy. If we moved to having a pass mark of 4000, and a first class boundary at 7000, it would be clear that we couldn't actually distinguish between a piece of work that got 6325 and another that got 6326. As numbers, 6325 and 6326 are different; they would, however, be translations of indistinguishable qualitative judgments. In a similar way, the marks 63 and 64 in our current marking system (for instance) are, arguably, arbitrarily distinct translations of the same qualitative judgment.

8. Our marking processes, however, generate an illusion of precision. When that a marker is faced with several scripts that they judge to be clustered in the 'low 2.1' bracket, they might give one piece of work 63 and another 64 in order to encode a sense that there is *some* qualitative distinction between them. Yet had the work been clustered differently, the marker might very well have given the first of these pieces a mark of 62 or a 64. We may be right that, when faced with two pieces of work at the same time, we can discern a fine-grained qualitative distinction between them. We are unlikely to be right if we think that, in general, we deploy the marks 63 and 64 in *reliably* distinct ways. That is: we are probably wrong to think that different markers, or the same marker on different occasions or in different contexts, will make use of these two marks in a consistent way. Research on the reliability of marking decisions in the humanities very strongly suggests that we do not.[1]

9. If our marking processes don't, in practice, yield a reliable distinction between, say, work given a mark of 63 and 64, any time and energy

[1] See, for instance, useful summaries in Sue Bloxham, 'Marking and moderation in the UK: false assumptions and wasted resources', *Assessment and Evaluation in Higher Education* 34.2 (2009), 209–20 and Sue Bloxham, Birgit den-Outer, Jane Hudson & Margaret Price, 'Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria', *Assessment & Evaluation in Higher Education* 41.3 (2016), 466–81.

spent by markers and moderators on distinguishing which of those marks to give in any given case is wasted. And any additional information apparently given to students by the difference between a mark of 63 and a mark of 64 on their work is an illusion.

10. One increasingly common response to this situation is to move to a marking scale that distinguishes fewer bands. If the above analysis is correct, such a move need not involve *any* loss of precision in our marking. It is not that such a move is a trade-off – some precision being lost for the sake of greater efficiency, say. No real precision is lost if we stop distinguishing between a 63 and a 64, if it is true that markers were never *reliably* making that distinction in the first place.

11. A step marking scheme might allow for a more transparent alignment between the marks that we give, the qualitative marking criteria that we employ, and the kinds of judgment that markers might actually be able to make. The 'noise' in our marking system is reduced, making the structure of our real qualitative judgments more visible.

12. For an example of a scheme that employs far fewer bands than ours, consider the Common Grading Scale used by the University of Aberdeen. It involves a 22-band scale, with a pass at band 9. If we take, for instance, a piece of Level 6 work that is of 2.1 quality, the current marking scale in Durham asks us to distinguish between 10 possible marks – 60, 61, 62, 63, 64, 65, 66, 67, 68 or 69. The Aberdeen scale distinguishes between three possibilities: 15, 16, 17 – effectively distinguishing between strong, average and weak 2.1s.

13. Durham University systems are set up to handle marks expressed as percentages, so we could not easily adopt a scheme like Aberdeen's. It is possible, however, to place restrictions on which percentages we actually use – and to use a so-called 'step marking' scheme within current systems and regulations. Such a scheme has recently been introduced, for instance, by King's College London, which has introduced a 24-band scale (see p.9 of their new 'Marking Framework'), with the possibility for some low-stakes assessments to use only a subset of these 24 bands. Keele University, to give another example, has a 21-band scale.

14. There are various ways in which such a scheme could be implemented for Common Awards. Take the same example of work which, at Level 6, is deemed to be of 2.1 quality. Our qualitative marking criteria

distinguish two categories of such work: 60–64 ('good') and 65–69 ('very good'). One fairly conservative way of implementing a step marking scheme would assume that markers can fairly reliably discriminate between work in the upper and lower halves of each of these two categories. We could therefore restrict markers to using only the marks 61, 63, 66 and 68 for work of 2.1 quality – cutting down from 10 bands in this area to 4 (and making similar restrictions for other segments of the scale).

15. A more radical approach would align our marks entirely with the qualitative criteria, and allow only one band per category – say, restricting marks in the 60s to 62 or 67.

16. We have modelled the impact that various such step marking schemes would have had on past students. In the spreadsheet that accompanies this paper, we have taken anonymised data on 125 BA students, 15 Diploma students, and 18 Certificate students, all of whom completed Common Awards programmes in recent years. The spreadsheet allows one to select a step marking scheme, and to see the impact on all the students' final AMWs.[2] The spreadsheet shows which AMWs increase and which decrease, and highlights any examples where the AMW increases or decreases by more than 0.5%,; it also notes any cases where an AMW is taken over a classification boundary, or taken into or out of a zone of discretion.[3]

17. Consider the fairly conservative step marking scheme described in §13 above, that allows marks of 61, 63, 66 and 68 (and similar options in every decade from the 30s to the 90s), and where we assume that 62s and 67s (and similar) would have moved up rather than down. Of the 125 BA students:

    a. 113 AMWs would increase and 12 would decrease;

    b. 15 AMWs would increase by more than 0.5%, with the greatest increase at 0.72%;

[2] This inevitably involves some guesswork. If a scheme allowed marks of 61 and 63, say, then it is not obvious which way a marker would have jumped for a piece of work to which on our current scheme they gave 62. The spreadsheet allows the modelling of different guesses.

[3] The AMW is the Arithmetic Mean Weighted mark – the average of a students module marks, weighted according to level. Students whose AMW fall within 2% below a classification boundary are in the 'zone of discretion', and the Board of Examiners may grant the higher classification if certain criteria are met.

c. no AMW would decrease by more than 0.5%, with the greatest decrease being 0.18%;

d. 4 students would be taken above a grade boundary, none below;

e. 1 student would be raised into a zone of discretion, and none would be lowered out of one.

The effects on Diploma and Certificate students are similar in scale.

18. Consider the more radical scheme described above that allows only 62 and 67 (and similar options in other decades), and where we assume that all work in the 60–64 bracket would get 62, all in the 65–60 bracket would get 67. Of the 125 BA students:

    a. 77 AMWs would increase and 48 would decrease;

    b. 26 AMWs would increase by more than 0.5%, with the highest increase at 1.23%;

    c. 7 AMWs would decrease by more than 0.5%, with the greatest decrease at 0.8%

    d. 5 students would have been taken above a grade boundary, 2 below;

    e. 1 student would be raised into a zone of discretion, 2 would fall out of such a zone.

19. Note that there are two kinds of information to draw from these results. One is to see them as telling us how much of a difference the change to step marking is likely to make to outcomes. The changes are not large, but they do differ in scale from scheme to scheme. This might help select a scheme that is likely to lead neither to dramatic grade inflation nor to dramatic deflation, and that is not likely to have a significant systematic effect on outcomes.

20. The other lesson to draw from these results is to see them as indicating the cumulative impact of the arbitrary distinctions that we make in our current marking practice. If markers can't reliably make finer-grained distinctions than those involved in the step schemes described above, then the difference in outcomes between our current schemes and

these step schemes don't reflect any greater precision or fairness in our current practice: they simply show us how the arbitrariness in our individual judgments does not always cancel out over the dozens of marked pieces that contribute to a student's final classification.

## The case against step marking

21. There are also arguments against a move to a step-marking scheme. First, the gain in efficiency might be more apparent than real. The idea is that markers avoid wasting time and energy deciding between marks that we can't reliably distinguish. Yet a marker who knows that they have to make a choice between (say) 63 and 66, rather than being able to use 64 or 65, may spend more time on this decision because it is more clearly consequential (and this effect is likely to increase the wider the steps are apart in the scheme chosen).

22. Second, although there may be a gain in transparency (with marks relating more simply to the bands in our marking scheme), that can be offset by the apparent artificiality of the scheme. The very fact that we will still, necessarily, use percentages might lead some markers and students to feel that the scheme represents an arbitrary restriction on the judgments that can be made, and so to feel that the mark given is not the 'real' mark. This impression may not be entirely alleviated by the arguments about reliability and precision made above.

23. Third, the move to such a scheme would involve a noticeable amount of work, both on the part of the Common Awards team (writing policy documents, producing training materials, and so on) and markers (who would have to learn knew habits). If it is not obvious that the introduction of the scheme would be solving a pressing problem, nor that it would bring very clear benefits, it may not be worth the work involved.

## Consultation

24. The arguments for and against the introduction of step marking seemed to the Common Awards Management Board to be fairly evenly balanced. We therefore want to discover whether there is widespread interest in such a scheme amongst Common Awards staff and students.

25. We would welcome responses from TEI Management Committees, indicating

a. whether you support the idea of moving to a step-marking scheme;

b. whether, if you do support such a move, you have a preference for the specific scheme you would like to see implicated; and

c. whether you think there are arguments for or against such a move that we have overlooked or misrepresented.

We would be grateful if such feedback could include opinions from student as well as staff members of the Committees.

Mike Higton
5 December 2023