Guidance on the scaling of marks for assessments

Introduction

1. This document summarizes guidance concerning the scaling of marks for assessments. The guidance arises from discussions and decisions at meetings of the University Learning, Teaching & Quality Committee (ULTQC), most significantly on 25 September 2012 but also as subsequently clarified.

2. It also provides additional guidance for Unit Boards of Examiners on treatment of:

a) borderline marks. and
b) structural mitigating circumstances (SMCs), by means of a formal process (see Annex 2) for the flagging of unit results where SMCs have not been able to be resolved through remedial action (such as scaling) or corrective judgement, such that the Boards of Examiners for Programmes (BEPs) may exercise limited discretion when determining students' final awards and degree classifications.

Principles

3. The responsibilities of the Board of Examiner for Units are set out in section 6.3 of QA35. Accordingly, its responsibilities include:

a. "ensuring the conduct of all examinations and assessments required to determine whether or not a student has successfully achieved the learning outcomes of the units under their academic authority";

b. "ensuring the academic standards of the units under its academic authority";

c. "ensuring that the summative assessments for a unit provide an appropriate level of academic challenge in testing that the learning outcomes have been achieved";

d. "determining the marks achieved by students taking units under its academic authority";

e. "ensuring that the finalised marks for individual units are an accurate reflection of the standards achieved by the candidates".

4. Thus the design context (**a**) is linked to the academic standards expected (**b**), with a matching level of assessment challenge (**c**) (e.g., Honours, Masters), such that marks can be determined (**d**), in a way that is properly calibrated to match the standards achieved (**e**). In general terms, therefore, *it would be reasonable to expect that well-established units at a certain level would elicit broadly similar results from a cohort of students*.

5. Scaling should, therefore, be limited to circumstances where it is necessary to deal with problems — that should only occur infrequently — in the relationship between marks initially recorded and the aim to ensure "that the finalised marks for individual units are an accurate reflection of the standards achieved by the candidates" (see para. 3.e above). Examples of such exceptional circumstances might include: the scaling of an entire unit's results if it had not been assessed on a basis comparable to that of other units; the scaling of the results for one group taking an assessment against that of another group if the first group's opportunity had been abnormally different (such as, a fire alarm in one examination venue but not in another where the same examination was being taken).

6. ULTQC's recent monitoring of scaling indicates that less than 3% of the University's unit marks are scaled (ULTQC minute 1187).

Analysis and scenarios

7. It is because the unit design and assessment contexts are well known in advance that scaling should not normally be required and should only be needed exceptionally.

8. In some disciplines, answers/essays are marked holistically and examiners can take a direct view of the range of marks (or degree class) in which a candidate's answer/essay belongs.

a. In these areas, criteria might be set out in advance to illustrate what would be demonstrated by particular levels of performance, and the examiners would assign marks within the spectrum associated with the level of performance they judge to have been achieved by each candidate.

b. Here, the "level" of the assessment is determined as much in the marking process as in the nature of the questions set, and scaling is less likely to be needed to compensate for a "difficult paper".

c. Scaling might still be considered in exceptional circumstances where, say, it appeared that the material covered in lectures was inconsistent with the examination questions that were set.

9. In some disciplines, examination questions commonly comprise a sequence of parts with a pre-specified marking scheme and the total mark for a question is the sum of marks gained section by section.

a. In ideal circumstances, a unit's mark scheme would be *designed* to deliver results that accurately reflect the standards achieved by its candidates. In a simple case, there might be eight key questions which would deliver a mark of 40% if they were all answered fully and correctly, such that further accurate answers beyond this would differentiate those who attained higher standards above the threshold pass mark. If all of the key questions were not answered fully and correctly, varying degrees of failure might be indicated. Getting seven of the eight questions right (at 5% each) could deliver a condonable fail mark of 35%.

b. In practice, there might be different ways of accruing a mark of 40%, and it might be important to know whether *all* sums of marks amounting to 40% equally indicated a threshold pass standard.

c. Simple calculations of aggregated marks might also need further consideration:

- In some cases, a mark scheme would not necessarily deliver an integer mark to be entered into SAMIS as the record of the student's achievement in the unit. This could arise if, for example:

  Fractions of marks were awarded within the mark scheme (e.g., five marks for a right answer, and half-marks for half-right);
  Each of three required answers were to be marked out of twenty, with the whole then turned into a percentage mark;
  Two components of a unit's assessments resulted in a composite mark overall, perhaps as an average of the separate marks;
  Two markers of the same project unit assessed it respectively as a marginal pass (42%) and a marginal fail (37%), with initial practice being to see what an average of the two marks suggested.

- If a mark scheme is so carefully designed that it is known an overall percentage mark of 39.5% will always be judged a fail, then such marks will always be rounded

down. However, if the status of such a mark could not be deemed the result of a carefully planned *automatic* process, then there must be a formal decision about the standard achieved and whether it is a pass (40%) or a fail (39%).

d. Ultimately, determining the correct integer marks to "ensure that the finalised marks for individual units are an accurate reflection of the standards achieved by the candidates" is not a scaling matter, but is the one of the proper concerns of the Board of Examiners for Units. All that has been said above about the pass/fail boundary is also true in other areas:

- The same goes for all units that are not *designated essential units* at the 34%-35% border for where the limit of condonability lies.
- Some units also have critical borders if they will count towards classification criteria. For example, Part 3 units in first-degree programmes are like this, and it means that the 49%-50%, 59%-60%, and 69%-70% borders are all significant.

e. However, examiners might also find that students have achieved marks on a paper which, taken together over the whole class, appear unusually high or unusually low. Setting a question and a marking scheme that will automatically deliver appropriate-looking marks is not always easy: scaling after seeing the students' performance allows a correction to be made for questions and marking scheme being too harsh or too lenient.

10. The distribution of marks can, of course, be affected by such factors as a small cohort size for a unit (*e.g.*, <10), but this does not mean that over-arching statistical monitoring across the units taken within a year, or across units taken over several years, is not worthwhile. The University has a policy (see QA35 para 6.5) for seeking to review and comment on results that, for example, appear to lie outside what might be expected, such that either a scaling action will be reported or its absence will be explained. Taking such information forward into the Annual Monitoring of Units can help to ensure that unit marks which required scaling in one year will not need to be scaled in the next. Boards of Studies are expected to monitor annually instances of scaling and those units whose mean marks fall outside the typical range.

11. The critical issue is always to "ensure that the finalised marks for individual units are an accurate reflection of the standards achieved by the candidates".

a. When reviewing average marks that lie outside a typically expected range, this does *not* mean, for example, that a higher average should simply be scaled to the point where it lies just inside the range deemed to be typical.

b. Unit marks should only be scaled if there is evidence that the marks initially recorded do not accurately reflect the standards achieved by the candidates, and any scaling that is required should move the marks to the range that would accurately reflect the standards achieved.

c. Accurately reflecting the standards achieved by the candidates depends not on a *norm-referenced* assumption that a set percentage of students should receive first- class marks (either for a single unit or overall for a degree), but on assessing performance against the *criteria* recognized as indicating such achievement.

d. Monitoring the relationship between marks and perceived standards of achievement in individual units will inevitably raise the question as to whether typical ranges of performance will be encountered in all cases, and what the causes of any differences might be. Since the aim is *not* to reduce all ranges of performance to match a typical range, reasoned argument must then be applied to establishing whether a set of results

for a unit should be scaled (and how far), or whether the difference now being encountered is produced by a group that is more or less able than might usually be observed — sometimes influenced by the small size of a group. In the former case, the action to scale should be matched by efforts to ensure that scaling will not be needed for the same unit for similar reasons subsequently; in the latter, the report of the higher or lower ability, and/or of the small size of the group, will serve to explain the difference in the particular case under consideration.

e. A final scenario can be used to emphasize the need for careful consideration of the evidence and for careful judgement in the reaching of a conclusion. If all students in a programme year take eight out of ten units while the remaining two unit slots can be filled with a variety of options, it might be expected that the eight units would tend to show similar ranges of performance from similar cohorts over time. However, it might be that some of the optional units could show, at least occasionally, much higher or much lower results than would be encountered across the eight units. This might be a consequence of very small numbers of students in some units, or it could be that two larger options tended to be selected by, on the one hand, the students who were at the less able end of the whole spectrum, and on the other hand, at the more able end. In this case, it would be mistaken to scale marks on both optional units to represent the same range of marks as would be encountered across the eight common units. The different results would be justified by showing that the students involved in each were representative of different levels of performance. It is in cases such as this that the ability to compare performance in one unit against the same students' performance across a range of units (by means of a scatter plot as illustrated in the Type 2 scaling example in Annex 1).

Key requirements for scaling

12. The process and conditions for scaling are set out in QA35 paras 6.5 and 6.6.

Supporting tools for scaling, and flagging structural mitigating circumstances (SMCs) as a measure of last resort

13. Advice on methods of scaling, commissioned from Professor Jennison is set out in Annex 1, Scaling methods.

14. Guidance on the flagging of SMCs which have not been able to be resolved through remedial action (such as scaling) or corrective judgement is provided in Annex 2.

Annex 1

Scaling methods

The possible approaches can be categorised into two types.

Type 1. Scaling one examination relative to general expectations

Here, the set of marks is considered and a judgement is made as to whether the marks are about right or in need of adjustment. Criteria could include the average mark gained by previous cohorts (should be around 55%, say), the proportion of students with a failing mark (ought not to be too high, based on normal experience), or the numbers gaining a first class mark (should not be excessively high, compared with normal experience). The scaling provisions refer to the use of historic information, specifically the set of unit marks over a three-year period, in making such judgements. Allowance must be made for random variation, particularly when only a small number of students take a paper.

Type 1 (a): A simple adjustment is to add or subtract a number to the marks of all students, *e.g.*, add 5% to all students for a paper with results deemed to be low (truncating marks to a maximum of 100% if necessary). Although this method looks simplistic, it can be surprisingly effective in correcting an anomaly.

Type 1 (b): A more sophisticated adjustment is to convert mark x to a revised mark $ax+b$, *e.g.*, use the conversion

$$\textit{revised mark} = 0.8\,x + 10$$

to raise low marks and reduce very high marks, leaving a mark of 50 unchanged.

Type 2. Scaling an examination against other related papers

A problem with scaling a paper in isolation is that it can penalise an exceptionally able cohort of students or be over-generous to a weak cohort. Statistical evidence of the need to scale is more clear-cut when students are seen to have fared very differently in one particular examination. Evaluation of such evidence is best done in a meeting of lecturers/convenors for a set of related units, *e.g.*, the set of units for a whole year or semester of a degree programme.

A check can be made by looking at the list of averages for units in a given year or semester. If one of these is out of line, consideration should be given to a scaling adjustment. Attention should be paid to any unusual or mitigating circumstances that may have affected the mark distribution. The scaling provisions mention problems in style or delivery of a unit. In discussing a possible downwards scaling, one should consider whether unusually good performance is due to well-motivated students working particularly hard on this unit, so the marks are deserved and do not need to be scaled.

The scaling provisions refer to a review of unit marks across the programme for a cohort of students. One way of doing this is the following. In assessing marks for unit A, plot a graph of each student's mark on unit A against that student's average mark on all other units (possibly excluding units in quite separate subject areas where one might expect discrepancies to be likely). If the difficulty of this examination is in keeping with others, data points will sit nicely around the 45º line, $y = x$, and most will stay within the pair of "tramlines" $y = x–10$ and $y = x+10$. If, for example, marks are consistently below the central line $y = x$ with a high proportion below the lower tramline, there is evidence that this exam was harder than the others. Scaling does not necessarily have to bring marks into complete alignment with the averages on other units; moving

half this amount could be deemed a balanced decision.

Conclusion

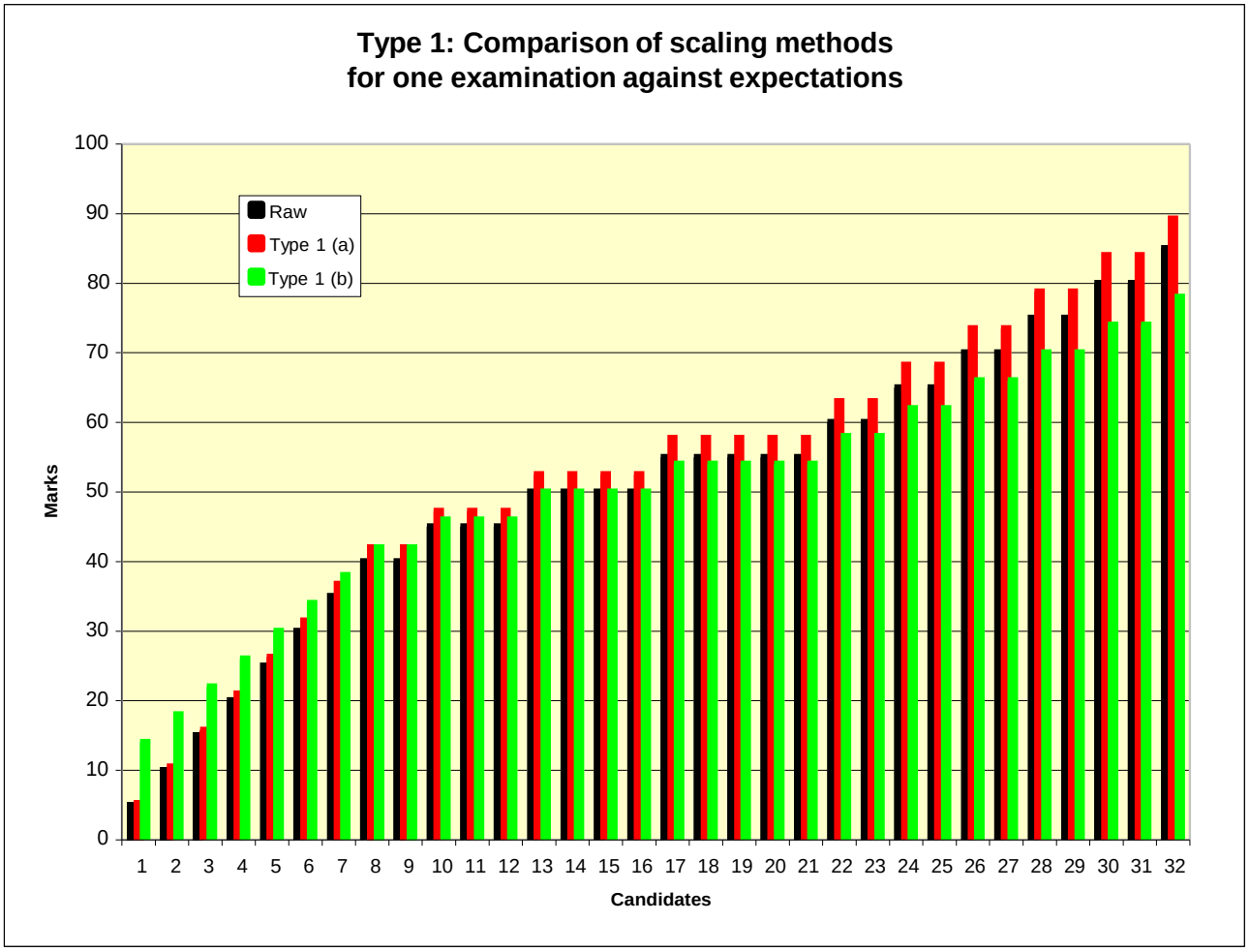
The overall intention of scaling is to be fair to the students, allowing a *post hoc* correction for an examination paper that has proved too easy or too hard.

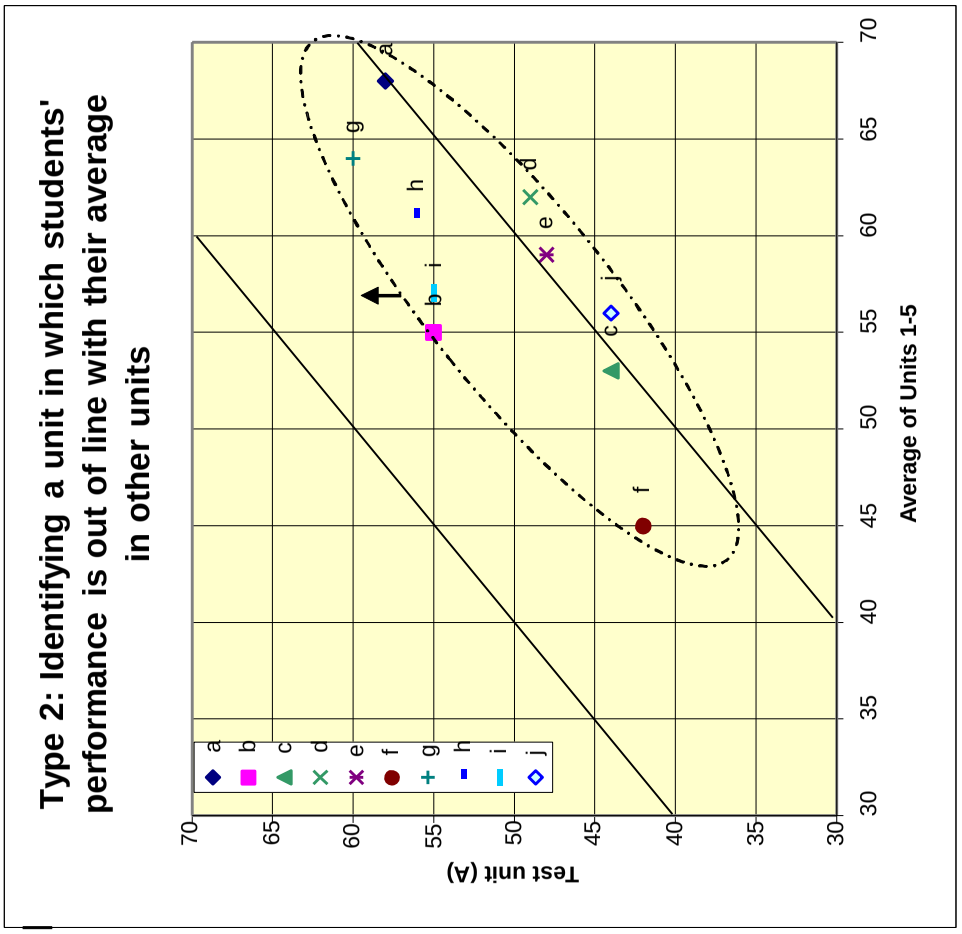
Evidence in the marks alone is confounded with other factors and scaling should only be adopted in exceptional circumstances where it is absolutely clear that it is needed.

CJ, updated September 2019 (no change to text).

[Worked examples and graphs are provided below.]

		Type 1 (a)	Type 1 (b)		
		Simple	More sophisticated ax+b		
		plus/minus %	a	b	ax+b
	Raw marks	5	0.8	10	
	Raw	Type 1 (a)			Type 1 (b)
	5	5.25	4	10	14
	10	10.50	8	10	18
	15	15.75	12	10	22
	20	21.00	16	10	26
	25	26.25	20	10	30
	30	31.50	24	10	34
	35	36.75	28	10	38
	40	42.00	32	10	42
	40	42.00	32	10	42
	45	47.25	36	10	46
	45	47.25	36	10	46
	45	47.25	36	10	46
	50	52.50	40	10	50
	50	52.50	40	10	50
	50	52.50	40	10	50
	50	52.50	40	10	50
	55	57.75	44	10	54
	55	57.75	44	10	54
	55	57.75	44	10	54
	55	57.75	44	10	54
	55	57.75	44	10	54
	60	63.00	48	10	58
	60	63.00	48	10	58
	65	68.25	52	10	62
	65	68.25	52	10	62
	70	73.50	56	10	66
	70	73.50	56	10	66
	75	78.75	60	10	70
	75	78.75	60	10	70
	80	84.00	64	10	74
	80	84.00	64	10	74
	85	89.25	68	10	78
Mean	50.47	52.99			50.38
Mode	55.00	57.75			54.00
Median	52.50	55.13			52.00





Student	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Average 1-5	Test unit (A)
a	65	65	75	70	65	68	58
b	55	50	55	70	45	55	55
c	60	55	40	55	55	53	44
d	60	60	60	65	65	62	49
e	55	55	65	65	55	59	48
f	40	40	45	50	50	45	42
g	55	65	75	70	55	64	60
h	55	55	65	65	65	61	56
i	50	50	65	65	55	57	55
j	45	50	65	65	55	56	44

This page is intentionally blank

Annex 2

Flagging structural mitigating circumstances (SMCs) as a measure of last resort

1. The following procedures provide a mechanism for the flagging of structural mitigating circumstances (SMCs) against students' records of achievement in a unit as a measure of last resort, used only where thorough investigation and deliberation have shown that it is **not possible** to otherwise ensure that the "finalised marks [...] are an accurate reflection of the standards achieved by the candidates". Structural circumstances may arise, very infrequently, where there is **no** appropriate remedial action that can be taken or corrective judgement that can be definitively applied. Defining and formalising the concept of flagging these most extreme types of SMCs provides clear guidance and a common process to ensure the fair and consistent treatment of students across the University. For these very infrequent occurrences, the Board(s) of Examiners for Programmes (BEP(s)) will subsequently note the problem and may exercise discretion regarding the effects of these structural circumstances when making judgements about final awards and degree classifications.

This decision to provide guidance was agreed by ULTQC on 8 July 2014. Minute 547 refers.

2. It is expected that, for the significant majority of these rare situations, remedial action will be taken or a corrective judgement applied. This will be in accordance with examples given in the *Individual Mitigating Circumstances & Assessment* (IMCA) document:

From time to time, a structural problem will occur with an assessment. For example, if a fire alarm disrupts an examination taking place in one venue but does not disrupt students taking the same examination in another venue, the Board of Examiners for Units should take appropriate steps to ensure that the results reflect common standards for all candidates. If something were to go wrong with one component of the assessment for a unit, but the rest was valid and those results could be relied upon alone, the Board of Examiners for Units would consider how best to judge the standards of performances achieved on the basis of the good evidence available. (www.bath.ac.uk/registry/imc/documents/imca.pdf, para. 4.e.)
or in the *New Framework for Assessment* (NFA) documents in relation to scaling:

2. (d) Consideration of any unusual or structural mitigating circumstances that might have contributed to a significant change to the mark distribution (e.g., a change in lecturer, particular acknowledged problems with a particular question or questions on an examination paper, recorded complaints from students about the style or delivery of a particular unit). (https://www.bath.ac.uk/corporate-information/new-framework-for-assessment/, Appendix 8)

3. It must be **demonstrated** that there is **no other remedy** through which to reach viable assessment outcomes, such as re-running the assessment, setting a replacement assessment, an appropriate use of scaling of unit marks (as indicated in the *Guidance on the scaling of marks for assessments*), or other means determined after consulting the Director of Academic Registry.

4. Significant concerns regarding the assessment of a unit may be identified prior to a meeting of the Board of Examiners for Units (BEU). These concerns should be immediately raised with the Director of Studies, the Head of Department, the Dean of the Faculty/School, and the Director of Academic Registry, in order that timely investigation and consultation might remedy the situation or, if no remedial action is deemed possible at that time, provide the BEU with the fullest evidence available to guide its decision-making. (Should such a problem be identified at a later stage, the BEU might have to be reopened.)

5. Should the BEU determine that it is not be possible to make an accurate judgement regarding the performance of the candidates within the unit based on the available evidence, it should formally record its concerns and recommendations in the minutes of the meeting.

6. The Dean of the Faculty/School, after consultation with the Director of Studies, the Head of Department, and the Director of Academic Registry, may then recommend to the Chair of ULTQC that the relevant unit results be flagged as affected by structural mitigating circumstances (the normal IMC-related "M" but associated with an appropriate SMC descriptive note) for each student.

7. The final decision regarding the suitability of applying the SMCs flag to a unit will be taken by the Chair of ULTQC. Where use of the flag is agreed, it must be applied to the unit results for the entire group of candidates taking the unit. The presence of structural mitigation on the unit outcome will be flagged within the Student & Applicant Management Information System (SAMIS) and on appropriate documentation as described in para. 6. Should the Chair of ULTQC not agree the use of the SMC flag, the Director of Academic Registry will advise the Director(s) of Studies of affected programme(s) and the Chair(s) of the BEP(s) of appropriate action. It will be the responsibility of the Director(s) of Studies to communicate the decision and the reasons behind it to the students affected.

8. The presence of structural mitigation on a unit outcome will be subsequently considered by the BEP when it reviews an individual student's overall achievement in the context of determining an award and a classification or grade. This consideration will be in line with the discretionary principles and processes already available for IMCs, separately described for first-degree and for postgraduate taught programmes. As with consideration of the effects of IMCs on academic achievement, BEPs should examine the fullest available evidence and statistical projections to determine the most appropriate outcome for the students. Existing spreadsheets from the IMCA may be used for these calculations.

9. Advice on the principles and procedures described here should be sought from the Director of Academic Registry in the first instance.

10. Agreement to use the SMCs flag will be reported to ULTQC as Chair's action. The Chair of ULTQC, in consultation with the Director of Academic Registry, will report to ULTQC after each academic year on any points that could contribute to future quality assurance and/or risk mitigation.

Annex 3

Principles for communicating BEU decisions about scaling

1. The aim of such communication is to aid student understanding about academic decision-making, thereby providing reassurance that scaling is undertaken fairly and robustly.

2. Any message will contain information about the decision as well as explanatory context and will adhere to the following principles.

a. Clarity

- Communicate in writing, and in plain English. Where specialist language/jargon is required provide a link to an explanation of the term.
- When required, refer students to existing documentation to avoid paraphrasing regulations or similar.

b. Authority

- Confirm an evidence-based decision made by a Board of Examiners exercising collective, academic judgement.
- Include an appropriate academic contact for further questions, normally the Director of Studies.

c. Context

- Acknowledge if students would have seen different, provisional, marks.
- Provide information on safeguards to ensure fairness and transparency, including institutional oversight.

d. Content

- Using the relevant BEU minutes, describe the process and reason that lead to agreement that the marks initially recorded needed adjustment to ensure *"that the finalised marks for individual units are an accurate reflection of the standards achieved by the candidates"* (QA35, Appendix 4, para. 5).
- Describe the method, avoiding 'before and after' marks.