

Policy on scaling of exam results

May 2015

This document outlines the School's policy and procedures relating to scaling of assessment marks; it is required by and expands on paragraphs 5.6–5.24 in the Assessment Handbook.

Scaling here refers to the systematic adjustment of marks across a cohort for an item of assessment (typically an exam), as distinct from *moderation*, which seeks to ensure that assessment standards have been applied appropriately. The School adheres to the standards for moderation outlined in the Assessment Handbook.

Scaling takes place when assessment marks fall outside typical distributions and where reasons can be found for this deviation which suggest that the marks do not accurately reflect achievement relative to academic standards, so that a realignment of marks is appropriate. Such factors may include mistakes retrospectively discovered in exam papers, particular questions turning out to be easier or harder than anticipated, or disruptions in the exams.

Note that a deviation from expectation might not occur across the whole range of marks; for example, it may be that surprisingly many candidates scored very high marks (owing to the hardest parts of the assessment not being hard enough), while the proportions of mid-to-low marks fall within expectations. The piecewise-linear scaling traditionally used in the School (and described below) can easily address such local deviations.

The need for scaling

Scaling is needed to ensure fairness across modules, with the intention that a given mark obtained in a module of a given level represents roughly the same level of achievement regardless of the module in which it was obtained. The marks obtained in mathematics exams tend to be more volatile than in other subjects, because they tend to be aggregates of lots of small parts with a quite precise mark scheme. This leads to unpredictable results, and can make it difficult even for experienced examiners to set a paper that will reflect students' achievement accurately. In addition, mathematics exams tend to produce a broader spread of marks than those in less rigid subjects – marks below 10 or over 90 are not uncommon – so that (even for a 'fair' exam which gives a greater reward to the more able candidates) scaling may be needed to ensure that marks correctly reflect candidates' attainment relative to the college scale. Scaling is therefore expected always to be a component of the assessment process in the School.

Procedures

Once exam marks have been recorded (either in SIS or in an in-house database) and checked, the first examiner should consider any summary data available, ideally including the proportion of candidates obtaining each grade, and compare them with typical distributions. (S)he should then make a recommendation to the exams team (through the Chair or Deputy Chair of the SEB) on whether scaling is required, and if so, what scaling should be used. The first examiner should re-examine scripts close to the proposed borderlines (especially the E/F borderline) and judge whether these scripts merit the proposed grades (even if no scaling is proposed). The first examiner may also use (and should communicate to the exams team) any other information relevant to the scaling decision, in particular a retrospective view on the difficulty of the assessment, or any problems associated with the assessment (such as an error in the exam paper). Any such information available should be taken into account, even if it is not available for all modules.

The exams team will consider the distribution of marks, the proposed scalings and any other information provided by the first examiner and make a final decision on scaling for the assessment. If

any scaling is applied, the exams team will carry this out (via the exams database or by downloading marks from SIS, modifying them and re-uploading the modified marks). Decisions should be recorded (with an explanation in the case where scaling has been applied), and a summary should be provided to external examiners ahead of the main SEB meeting and also included in the paperwork for the SEB meeting.

Basic principles

1. Scaling should not be purely normative: we do not scale simply because (for example) the proportion of A grades is out of line. We use such deviations as an indicator that scaling *may be* necessary, and examine whether there is an academic justification for scaling. Academic reasons for scaling fall into two categories:
 - problems with the assessment, e.g. a retrospective view that the exam was too easy or too difficult, an incident during the exam, or inappropriate standards of marking;
 - problems with teaching and learning, e.g. an over-ambitious syllabus, an inexperienced lecturer, or poor student engagement.

In either case steps should be taken to improve the situation for the future, in addition to any scaling for the present cohort.

2. Under-scaling is better than over-scaling, especially when scaling marks down.
3. Scaling should be applied consistently to all candidates for a particular item of assessment, including re-sit candidates.
4. Scaling should be applied to individual assessment components, not to overall module marks. Only components worth 30% or more of the module mark can be scaled. Note that 'components' here means the elements considered as components by SIS, even though these may themselves be aggregates (e.g. of weekly coursework).
5. Scaling should only very rarely be applied to small cohorts (fifteen or fewer students).

The scaling formula

Scaling should be carried out using the 'adjusting borderlines' method that has been used in the School in the recent past: we identify appropriate 'internal' grade boundaries, and apply the (simplest) piecewise-linear function mapping 0 to 0, 100 to 100 and the internal grade boundaries to the college grade boundaries. In particular, the scaling function should be an increasing bijection from $[0, 100]$ to $[0, 100]$.

This method should always be applied, even in cases where scaling is being applied because (for example) a particular question has been retrospectively judged to be too difficult; scaling by discounting just that question would require considerable extra work and would effectively penalise candidates who had succeeded with the difficult question.