The Decimal Grade System of Degree Classification: A Guide to its Principles and Development

Peter Millican, University of Leeds

The purpose of this document is to discuss the background and principles of the Decimal Grade System of degree classification now operating within the University of Leeds, and to explain its main features in terms of the very important constraints that it was designed to satisfy.

1. The New Context: Publicity, Modularity, Interdisciplinarity, Anonymity

When considering the advantages and disadvantages of different methods of degree classification for use in a modern British university, it must constantly be borne in mind that the context within which such methods must operate has changed radically over the last decade. Here it will be sufficient to draw attention to four of the most important changes:

(a) Publicity

All marks and classification procedures are now made public to students, with methods of calculation being announced in advance and results made available for scrutiny afterwards.

(b) Modularity

Within many universities, degree programmes are now entirely modular, enabling students to "mix and match" components rather than taking a pre-planned integrated programme.

(c) Interdisciplinarity

Many degree programmes are now designed to be cross-disciplinary, and even those that are not often give scope for interdisciplinary components.

(d) Anonymity

Student numbers imply relative centralisation and automation of degree classification, and there is increasing pressure towards anonymous classification on principle, to eliminate potential bias.

Publicity, which became inevitable in the wake of data protection legislation and quality audit requirements, has the consequence that inconsistent marking practices and unjust or anomalous methods of classification would be very likely to be discovered, giving a serious risk of increasing numbers of student appeals and public embarrassment to the university concerned. It also implies that classification policies are highly likely to influence student behaviour, so that certain traditional practices that are potentially exploitable by the "strategic student" (e.g. discounting of results in "electives", or of a student's worst few marks) become seriously problematic. *Modularity* and *interdisciplinarity* require that any standard classification procedure should be able to accommodate a wide range of marks, from modules of various sizes and structures (including large multiple component modules) and from a wide variety of disciplines in which marking practices have

traditionally been very different. They also doom to failure any attempt to divide the spectrum of interdisciplinary, joint, and major/minor degree programmes into coherent and distinct academic categories (e.g. "Arts" and "Science") that can then claim justifiably to be classified quite differently. *Anonymity* implies that classification procedures can no longer rely heavily on the checks and balances provided in the past by departmental examiners' meetings, in which the careful consideration of a student's overall performance, by those well acquainted with his or her work, could overrule if necessary an inappropriate decision of a mechanical degree classification algorithm.

We shall now see how these four changes combined to make the retention of traditional degree classification systems quite untenable within the University of Leeds. The introduction of the Decimal Grade System, therefore, was not merely a change for its own sake: a new system simply had to be chosen, and the detailed analysis summarised in the following pages indicated that of those available the Decimal Grade System was the one easily best suited to satisfying the new constraints.

2. The Incompatibility of Traditional Classification Systems

The long debate on degree classification within the University of Leeds (in the wake of its decision to "go modular") was provoked largely by the realisation that the various traditional methods of classification hitherto used in its different departments were not jointly compatible, and were probably not even individually suitable for the treatment of potentially interdisciplinary modular degrees. The most obvious divide was between those departments (principally in the Arts and Social Sciences) that relied on "class profiling" methods, and those (principally in Science and Engineering) that relied on "mark averaging" methods. This distinction had implications which extended well beyond the mechanics of classification, for it also had a profound effect on marking practices within the different departments. In a "class profiling" department the exact mark awarded to an examination script would typically be of little significance: all that mattered was the "classband" into which the mark fell (i.e. whether the mark was in the "First Class" or "2:1" range etc),[1] so that 60 would do as well as 69, and 70 as well as 90. Thus markers could confine their most painstaking consideration to borderline decisions (such as the choice between 69 and 70), being relatively relaxed elsewhere and limiting Fail and First Class marks to the 30's and 70's respectively – all this without any worry that BA students would thereby be treated unjustly by comparison with their scientific peers. By contrast in a strict "mark averaging" department the mark was everything, and the "classband" of no independent significance. This had contrasting effects in different parts of the scale: in the middle it avoided the hardest decisions, because for example the choice between a mark of 59 or 60 need be no more agonising than the choice between 64 or 65; but at the top and bottom, it gave a significance to the extreme marks far greater than they ever had in the "class profiling" departments – here it could make a real difference to the student's degree class whether the mark awarded was 10 or 30 for a bad Fail, or 75 or 95 for an excellent First.

[1] The term "classband" is preferable to "class" in this context since it enables the latter to be reserved for talk of *degree* classes and the corresponding qualitative standards as opposed to numerical ranges of the marking scale.

3. Other Objections to Traditional Classification Systems

Although the conflict between the "Arts" and "Science" traditions (and the need for any University-wide modular structure to incorporate both in a coherent manner) provided the most obvious imperative for reviewing the traditional systems of classification, this was not by any means the only reason for doing so. For both "class profiling" and "mark averaging" systems had their own independent difficulties which made them very unsuitable for the new situation in which, because of modularisation, degree programmes and even individual modules would have far greater flexibility of content than previously, and degree classification must inevitably in many cases take place relatively mechanically and anonymously.

The independent objections to the "class profiling" paradigm are particularly fundamental, since modularisation itself, but also in particular the increasing use of a variety of assessment components within individual modules, are very dubiously consistent with the practice of treating any mark in the same "classband" as identical for classification purposes.[2] The whole logic of modularisation presupposes that modules can properly be assessed in relative isolation, while the method most commonly used for deriving an overall mark on any particular module – namely the arithmetical averaging of component marks – extends this principle to sub-modular components. In such a context it simply becomes indefensible to treat a module mark which is little more than a conflated average as a result whose significance changes enormously between the values of, say, 59 and 60 – if a module mark of 59 is achieved as an average of 64 and 54 on two sub-modular components, and a module mark of 60 is achieved as an average of 64 and 56, then the real difference in significance between the 59 and the 60 is clearly no greater than the difference in significance between the marks of 54 and 56 on the second component.[3] It might be suspected that this sort of point, though theoretically undeniable, can in practice be ignored because it misrepresents the nature of marking and classification: neither is an exact science, it might be said, and the virtue of a classband-based system, for all that it can seem anomalous when examined in detail, is that it sensibly acknowledges the rough-and-ready approximation which characterises all such assessment. But such a response can be decisively refuted, for even if marking is indeed only approximate (as it surely is), this in no way justifies the gratuitous introduction of unnecessary *additional* error into a process as significant and sensitive as degree classification.[4]

---

[2] Quite apart from the *obvious injustice* of doing so when a student might have a whole run of marks that fall just below a classband borderline, and hence have a very significant element of achievement (well above the bare classband minimum) simply ignored for classification purposes on module after module. Other students, of course, will receive a corresponding benefit when they achieve marks that just cross the classband thresholds, and so the overall spread of degree class awards is still likely to look "sensible". But this cannot remove the suspicion that sometimes the classes are going to the wrong people. Some particularly striking examples of such historical injustices were highly influential in persuading the Leeds University Arts Faculty to move away from the "profile-based" classification procedures under which they had occurred.

[3] For a straightforward example of inconsistent classification arising from this sort of anomaly, see the appendix section entitled "The Classical Culture Problem".

[4] See the appendix section entitled "How Fine-Grained Should Our Marking/Classification Scales Be?", for a quantitative assessment of the additional error which use of a classband-based system introduces. Under the plausible assumption that examiners' perception is usually accurate to around plus or minus one fifth of a class, the use of a classband-based scale can be shown to magnify the risk of misclassification over the "background" rate by around 169% – making the risk more than two and a half times as great – and apparently for no good reason whatever.

Independent objections to the "mark averaging" paradigm are based less on pure logic than on the actual traditions of academic judgement and its representation within that paradigm. First, it is a matter of historical fact that even in the most committed mark-averaging departments, the classbands into which marks fall have in practice been accorded a far from negligible significance.[5] But this seems inconsistent with treating marks of, for example, 58, 60 and 62 as though they were for classification purposes simply evenly spaced points on a linear scale. A second objection relates to the extent of the traditional "percentage" scale, ranging from 0 to 100 with the pass/honours level around 40 and the First Class starting at 70. This has the consequence that the lowest Fail is a full four "standard classband widths" below the Third Class threshold, while a perfect result is three standard classband widths above the First Class threshold. So in a mark averaging system just a few of these extreme results can have what to most examiners seems a disproportionate effect on the degree class (e.g. twenty three marks of 71 and one of 0 average 68; twenty one marks of 66 and three of 98 average 70). The impact of such extreme marks might not seem so inappropriate if they were properly reserved for truly exceptional performances, but again it is a commonplace that actual practice here is far from consistent – indeed most very high and very low marks are assigned not on the basis of a judgement that the work done is out of the ordinary, but rather from the mechanical application of additive marking schemes, over a range of disciplines where the standard of performance needed to achieve any given percentage can vary dramatically. Though often taken for granted, it is surely *seriously anomalous* in a mark-averaging classification system that those assessments traditionally *most likely* to yield marks of 90 and above (e.g. examinations involving rote learning; multiple choice tests; exercises in basic mathematical or other formal techniques, or in elementary language translation) should in practice be precisely those which are *least suited* to eliciting, and identifying, a genuinely outstanding student performance.[6]

4. The Search for a Compromise Between Class Profiling and Mark Averaging

Those who worked within the class-profiling and mark-averaging traditions were often well aware of these limitations, and took them properly into account in examiners' meetings. Thus a student whose class profile seemed ungenerous in the light of a long sequence of marks near the top of a class would typically be treated charitably, as would one whose mark average was brought below a threshold by one extreme Fail (especially if a sufficient number of other marks were within the higher class). Such checks and balances, by which examiners within one tradition would incorporate elements of the other, were in most cases entirely informal, but as student

[5] For example it has been common practice, even in mark-averaging departments, to summarise students' performance for classification meetings in terms of the number of credits falling within each classband. Though perhaps originating merely from the technical difficulty of producing a fine-grained graphical representation of the spread of results (rather than any deliberate intention), this naturally biases examiners' attention towards classband-focused assessment.

[6] Part of the problem here is a fundamental lack of clarity about what exactly a "100%" result is supposed to mean – for example does it signify a piece of work: (a) in which all the answers are right irrespective of their difficulty?; or (b) which represents a performance as good as one could expect from anyone in the entire history of the human race?; or (c) which represents a performance as good as one could expect from the best student one is ever likely to teach? In "right or wrong" disciplines (a) is probably the usual choice, but this clearly provides no consistent indicator of performance quality, and thus sits very uneasily on a scale whose other principal points are supposed to be calibrated according to standard "classes" of performance such as "First Class" or "2:1 quality". In discursive disciplines (b) is perhaps most commonly assumed, which would explain why marks above 80 are so unusual, but it is surely most logical to take (c) rather than either of the others as the appropriate calibration for the highest point of an undergraduate marking scale.

numbers rose dramatically and the scope for such special consideration correspondingly reduced, it became increasingly important to formalise and automate them. Hence modularisation, and the consequent need to combine typical "Arts" and "Science" marks coherently and justly, was by no means the only pressure behind the search for an acceptable method of reconciling the "class profiling" and the "mark averaging" paradigms.

Various attempts to achieve such a reconciliation were proposed, including some which explicitly combined the two paradigms by allowing students to qualify for a given class in either of the two ways, or by offering a trade-off between the two (so that for example a better "class profile" would reduce the mark average classification thresholds). Another suggestion in much the same spirit was to take account of both the mark average (i.e. the *mean*) and the mark *median*, with the latter functioning in much the same way as the class profile to minimise the impact of extreme results. Though sometimes initially plausible, unfortunately none of these compromise solutions proved tenable: even the best of them yielded counterintuitive results in some circumstances, and serious logical or mathematical anomalies in others.[7] On reflection this is perhaps hardly surprising – when similar marks are used to represent different standards of achievement between different disciplines, or even between different skills within the same discipline (e.g. translation, formal logic or factual tests as opposed to literary, philosophical or scientific essays), no amount of purely arithmetical manipulation can be expected to iron out the resulting inconsistencies. A more principled approach to the problem was evidently required.

5. Resolving the Marking Scale (A): Linear Representation of "Judgemental" Marking

The first stage in the genesis of the Decimal Grade System was the development of a form of marking scale capable of representing student performance consistently, whether the marking concerned be done in the "judgemental" manner characteristic of the "Arts" tradition or in the "additive" manner characteristic of the "Science" tradition.[8] Given the ultimate aim of creating a straightforward and coherent classification system, it was obviously desirable that this scale should provide a *linear* representation of student quality of performance (i.e. with each constant increment on the scale representing a constant increment in quality), since only thus could simple averaging – an easily understood, widely applied, and provably coherent procedure – be legitimated as a method of combining such marks. Hence the first step in developing the scale was to determine what actual judgements of quality were implicit in these traditional marking practices, particularly in the judgemental marking of the "Arts" disciplines which tended to be highly non-linear near class boundaries (presupposing for example a far larger increment in quality from 58 to 60 than from 60 to 62). To address this issue a questionnaire was circulated to all departments in the Arts Faculty, including questions designed to elicit their implicit judgements regarding the relative significance of marks (as understood on their traditional marking

[7] For an explanation of some of these anomalies, see the section in the appendix entitled "Mathematical Anomalies in Degree Classification Systems". Such anomalies are not just a theoretical nuisance, since they go together with undesirable implications for student motivation (e.g. using the median mark for classification gives students a strong incentive to focus their energies on their favourite 130 credits, and to content themselves with merely passing the remaining 110 credits).

[8] See the appendix section entitled "Marking on the Module Grade Scale" for an account of this important distinction.

scale) and classes. The entire questionnaire and a summary of responses is given in the appendix,[9] but for present purposes the crucial outcome was a clear endorsement of what we might call the "Thick Borderline Principle":

> *As a representation of examiners' judgement concerning the quality of a student's performance, the classband into which a mark falls is significant, but nevertheless there is a more significant difference in quality between the top and bottom of the same classband than there is between the nearest points of adjacent classbands.*

This implied that in order to provide a faithful linear representation of such examiners' judgements, the desired scale must enforce a significant numerical difference between marks falling immediately on either side of each classband borderline (implying a "thick borderline"), while still allowing for an even greater difference between marks at the top and bottom of the same classband. It clearly made sense to retain, as far as possible, the traditional 10-point classbands to which markers had become accustomed, not only because most examiners were comfortable with this, but also because their habituation to it provided an important safeguard of common standards from year to year and across disciplines.[10] Within this traditional framework, the most natural and appropriate way of enshrining the Thick Borderline Principle seemed to be to require *marks that were intended to signify a definite judgement of class* to be confined within the seven-mark middle range of each decade (e.g. from 62 to 68), leaving the three cross-border marks (e.g. 59, 60 and 61) as explicitly borderline or classless. This thick borderline had the desired effect of forcing a 4-mark numerical distinction between marks falling *determinately* on either side of each classband boundary (e.g. 58 and 62), while leaving room for a greater, 6-mark numerical distinction between the top and the bottom of each individual classband (e.g. 62 and 68).

An important side-effect of enshrining the Thick Borderline Principle implicitly in this new "gappy" scale (rather than attempting to enforce respect for it by complicated explicit rules) was that all the mathematical anomalies which had so beset the class-profile systems could immediately be eliminated by moving to classification on the basis of mark averaging, but without losing the partial emphasis on classbands which remained attractive to judgemental markers. For now the "classwise" significance of a judgemental mark would be built into the mark numerically – instead of giving a mark of 60 to work of marginal 2:1 quality, the marker would give 62, so that the student would immediately, so to speak, receive a "bonus" for achieving a 2:1 score. And thereafter, if this mark were to be added or averaged with others, it would always carry that bonus with it, and no special rules or further manipulation would be required to ensure that the initial class judgement continued to carry its due weight. Thus all the old problems regarding conflated marks and mark averaging could be eliminated completely just by insisting that *marks intended to signify a determinate judgement of class*

---

[9] See the section entitled "Questionnaire on the Ideal Classification System for Joint Honours BA Degrees". The quoted principle is implicitly supported very strongly by the responses to questions 8, 9, 11 and 12.

[10] This is a very important consideration in the maintenance of standards, but one often overlooked in policy discussions. Marking is a process involving enormous numbers of examiners across numerous disciplines, so any major disruption to marking procedures (such as moving to a marking scale that is *radically* different from what they are used to, especially over the "central" range where the vast majority of marks occur) is extremely costly in examiner time and stress. Indeed such a

should be given non-borderline values (e.g. avoiding 59, 60 and 61), but such borderline avoidance applies only when marking judgementally, not in any subsequent arithmetic processing. If a mark is a mere conflate, a total, or a result of linear interpolation (as is common in additive or marking scheme assessment), or if the marker does not judge that the work concerned falls determinately within any particular class, then to ascribe that mark a class-based significance is obviously quite inappropriate. By building in the "class bonus" implicitly at the point of marking, rather than resorting to later classband-based manipulation of previously existing marks, the new system could completely avoid such inappropriate treatment. This offered an elegant solution to the problem of reconciling the judgemental marking characteristic of the Arts tradition with the additive marking and mark-averaging classification characteristic of the Sciences.

Marking in accordance with the "Thick Borderline Principle"

The Thick Borderline Principle was enshrined within the new marking scale by ruling that (within the main body of that scale) marks ending in "9", "0" or "1" should be treated as *borderline for the purposes of judgemental marking*. Thus an examiner who wishes to signify that an assessment falls determinately within a particular class should avoid giving such a borderline mark (hence to signify that an assessment is of determinate 2:1 quality, a mark between 62 and 68 should be given). Qualitative descriptors were defined making this distinction completely explicit – around the 2:2/2:1 borderline, for example, these descriptors are:

65-66	Middle 2:1	62	Marginal 2:1	57-58	High 2:2
63-64	Low 2:1	59-61	Borderline 2:1	55-56	Middle 2:2

Borderline marks accordingly correspond to a judgement that the assessment is of borderline (or mixed) quality. The distinction between borderline and non-borderline marks is irrelevant to non-judgemental marking (e.g. "additive marking", based on an additive marking scheme), and should play no role in subsequent marks processing (e.g. when averaging or conflating marks).

6. Resolving the Marking Scale (B): The Extent of the Module Grade Scale

With the central 40-70 range of the new marking scale (the "Module Grade Scale") now in place, it remained to be determined how the extremes of that scale should be organised, and in particular, what should be the width of the First Class and Fail ranges. Here a number of considerations came into play:

(a) Justice, consistency, and the use of averaging for combining marks all clearly required that the marks outside the central range, like those within it, should represent *agreed standards of quality of student performance on a linear scale* – this implied that they could not be defined in terms of a crude

disruption positively invites inconsistencies as so many individuals across the university, usually under great pressure, all struggle to reinterpret their traditional marking standards within the new scale.

"percentage of right answers", and hence there was absolutely no *prima facie* case (despite the common assumption to the contrary) for adopting the full 0-100 scale traditional within technical disciplines.[11]

(b) Analysis of past data indicated that when marking on the full 0-100 scale was available, fewer than 1% of marks awarded had been in the 1-24 range, with most of these being in the 20-24 range. Around 0·7% of marks were zeros (presumably for absence or non-submission), but apart from these fewer than one mark in 1000 fell below 10. Likewise at the other end of the scale, fewer than 1% of marks had been above 85, with most of these lying in the 86-90 range. Nearly all of the "extreme" marks, moreover, came from a very few departments (notably Physics, and Electrical Engineering), implying a serious potential injustice if mark averaging were to be used for degree classification in Joint Honours or other cross-disciplinary degree programmes.[12]

(c) If degree classification was to be based on mark averaging, then it was important to avoid the risk (explained in section 3 above) that extreme marks might have a seriously disproportionate effect on the class awarded. This suggested that the traditional 0-100 scale should be truncated at the top, and even more at the bottom (except in cases where a deliberately punitive mark, for example "0" for non-submission of work or uncondoned absence, was explicitly intended). On the other hand it was important to provide sufficient range in the scale to allow very good, and very poor, marks to have an appropriately signficant effect on the overall average – hence the traditional practice in some "Arts" disciplines, of using marks outside the narrow 35-75 range only for exceptional or appalling performance, would have to change.

When the faculties were consulted about all this, some initially preferred a narrower range of marks (e.g. 30-80) and some the traditional "percentage" scale (i.e. 0-100), but fortunately over the consultation period a clear consensus emerged for a marking scale ranging from 20 to 90, implying that the best First Class mark would be two "standard classband widths" above the First Class threshold, while the lowest normal Fail mark would likewise be two standard classband widths below the Third Class threshold. The top of the scale could thus be understood as implying a distinction (already implicitly recognised in the marking practices of many departments) between two levels of First Class performance – just as there is a division between "Lower Second Class" and "Upper Second Class", so there would now be an analogous division between (ordinary) "First Class" and "Excellent First Class" marks. Likewise at the bottom of the scale a distinction would be drawn between (ordinary) "Fail" and "Bad Fail" marks, with the latter being in the 20-29 range. However there was also clear support for recognising two explicitly punitive marks below the 20 minimum of the main scale, one

---

[11] The unreflective assumption that "marks are percentages" cannot stand up under critical scrutiny, as we shall see below, but it is nevertheless extremely widespread amongst "additive" markers, for it seems to be naturally suggested by the use of additive marking schemes.

[12] It would clearly be unjust if, for example, a student with an excellent First Class performance in Physics and a low 2:1 performance in Philosophy should be awarded a First Class degree in Physics and Philosophy, whilst a student whose performance was the mirror image of this (i.e. excellent First Class in Philosophy, low 2:1 in Physics) should achieve only a 2:1, simply because the mark used to represent an "excellent First" in Physics is high enough to pull up the average, whereas that used in Philosophy is not. This again emphasises how a mark averaging system fundamentally presupposes that the marks being averaged represent *consistent standards of quality of performance on a linear scale*.

being the universally understood zero mark for absence or non-submission of work, and the other an intermediate punitive mark (equivalent to 10 when averaging) indicating "no serious attempt", for cases where the quantity of work done on a module is deemed unworthy of a mark on the main scale. Such punitive marks are intended to target the increasingly common "strategic student", giving such students a clear disincentive to neglect any modules which they are supposed to be taking seriously.

Summary and Terminology: "Marks" and "Grades"

The "Module Grade Scale" was thus agreed to run from a maximum "grade" of 90 down through the "Excellent First", "First", "2:1", "2:2", "Third", "Fail" and "Bad Fail" classbands, to a minimum normal grade of 20, but with two special punitive grades below this, namely "NSA" (no serious attempt) and "0" (absence or non-submission). All of the numerical grades on this 20-90 scale are calibrated by qualitative descriptors, which define the quality of performance that they are intended to signify – these descriptors play an explicit role when marking judgementally (see previous box). Additive or other non-judgemental marking, as we shall see in Section 7, is calibrated less explicitly, but conformity to the same standards can be assured by reference to the classband boundaries (e.g. by ensuring that the quality of performance yielding a mark of 70 is indeed of borderline First Class standard), followed by interpolation of intermediate grades. The Module Grade Scale thus defines a consistent qualititative standard applicable to all types of marking. From now on in this discussion, the numerical values on this scale will be referred to as "grades", in order to distinguish these standardised values from ordinary "marks" which can be on any scale (e.g. "marks out of 20" or "marks out of 150", as well as "marks out of 100"), and may or may not be adjusted to conform to an agreed qualitative standard.

7. Resolving the Marking Scale (C): Additive Marking on a Quality-Calibrated Scale

Agreement to the 20-90 Module Grade Scale implied changes, at the extremes, to the typical marking practices of both "judgemental" and "additive" markers, raising the question of how consistency over such changes is to be assured.[13] In the case of judgemental marking the matter is relatively straightforward, because as we have seen above, the Module Grade Scale is explicitly intended to provide a (linear) representation of qualitative standards, and hence its various points are appropriately calibrated by judgemental descriptors such as "Middle 2:2" or "High 2:1". Accordingly a judgemental marker who deems the quality of a student's script to match one of these descriptors (for example "Marginal Excellent First") can be expected to assign the corresponding grade (here 82) even if this is different from the mark that might have been assigned under traditional "Arts" marking

[13] Though this should not be taken to imply that such consistency was in any way assured before the changes! Indeed it is clear from past marking statistics that the marking patterns in different departments (even in closely related discplines) sometimes differed considerably, especially at the top and bottom of the scale. So measures to encourage consistency, such as those described here, would have been highly desirable even if there had been no change whatever in the marking rules.

practices (perhaps 75 in this case). As long as all examiners pay heed to the descriptors and mark accordingly, their judgements should be given a consistent numeric representation.

Ensuring the consistency of additive marking is rather less straightforward because this form of marking is less direct than simple judgemental marking, being mediated by the use of an additive marking scheme. Here the marker will typically: (a) use a piecemeal marking scheme to assign a mark to each component of the work done; (b) add the component marks to yield a "raw total" for the work as a whole (often expressed as a percentage); and (c) award an overall grade to each student on the basis of the raw total achieved. It is important to recognise, but often overlooked, that this last stage is *not* automatic – it requires the marker to judge which raw total corresponds to each of the principal reference points on the appropriate qualitative scale, and to award grades accordingly. In principle an examiner may succeed in creating an "ideal" marking scheme, in which there is a fairly precise correspondence between raw totals and grades (so that, for example, a raw total of 40% indicates a student performance that just reaches Third Class quality and therefore merits a grade of 40, while a raw total of 70% indicates a performance that just reaches First Class quality and therefore merits a grade of 70). But although marking schemes are indeed often designed with this sort of correspondence in mind, there is no reason to suppose that it is always achievable over the entire grade range, even as an approximation.[14] This is not to say that expressing the results of an individual test in percentage terms is necessarily inappropriate, but if those results are to be combined with others in any sort of averaging or other arithmetic procedure (as they are in degree classification), it is a manifest requirement of fairness and justice that all these results should be represented on a common scale, whose appropriate calibration is in terms of quality of performance, not mere "percentage of right answers".[15]

Once it is recognised that some judgement and calculation is required to translate the raw totals produced by additive marking into standardised grades on the Module Grade Scale, the translation process itself is fairly straightforward. All it involves is the examiner's judgement as to which "raw totals" correspond to the principal reference points on the Scale (e.g. the Third Class and First Class boundaries), followed by an arithmetical process of interpolation. Suppose, for example, that the marking on some examination is done out of a possible total of 150, and the examiner judges that a raw total of 66/150 just reaches the Third Class boundary (grade 40) while a raw total of 109/150 just reaches the First Class Boundary (grade 70). Then the grade corresponding to a raw total of 92 may be calculated by linear interpolation as $40 + (92-66)\times(70-40)/(109-66) = 58$. Many

---

[14] To appreciate why, contrast one subject (e.g. perhaps the Philosophy of Kant) in which the initial marks are relatively easy to get but the later marks much more difficult, with another (e.g. perhaps Formal Logic) in which it is hard to get started, but where mastery of further material then follows relatively quickly. This kind of contrast seems in many cases to be intrinsic to the nature of the subjects themselves, and if so it may be impossible to devise marking schemes for examinations on the two subjects both of which yield a linear relationship between raw total and quality of performance. Indeed in the one case the graph of raw total against quality is likely to be convex, and in the other concave.

[15] Any remaining temptation to assimilate percentages and grades can be undermined by considering an examination in which even the common approximate correspondence between the two clearly breaks down – for example, an examination so difficult that even the very best students in a large and talented cohort score only around 50% of the marks available, or one so easy that the median score of a large but average cohort is around 75% of the marks available. Here it is *obvious* that the percentage marks cannot properly be taken as module grades, and this illustrates that in general there is an independent standard (i.e. the accepted scale of quality of performance in terms of "First Class", "2:1" etc) to which the percentage marks must conform if they are to be treated as reliable grades for classification purposes.

examinations are designed in such a way that all grades in the 40-70 range can be dealt with by this sort of simple linear interpolation between the 40- and 70-points, but outside that range (and also sometimes within it, if moderation and inspection of the scripts indicates that linearity has not been achieved) it is likely that additional reference points will be needed. All such reference points may be taken straightforwardly from the qualitative descriptors of the Module Grade Scale (thus for example the grade of 80 should be given when the raw total indicates a borderline "Excellent First" performance, one standard class width above the 70-point). In this way the very same qualitative scale which ensures the consistency of judgemental marking also ensures the matching consistency of additive marking, albeit now through the intermediary of the marking scheme. Hence after all this processing, the results of judgemental marking and the results of additive marking can be consistently combined together.

8. Degree Classification by Grade Averaging, and the Module Weighting Issue

With the development of the Module Grade Scale, it became possible to average module grades together without fear of distortion, irrespective of whether these had been generated by additive or judgemental marking. This therefore paved the way for degree classification based on the results of such averaging, but raised the important issue of which module grades should be included in this average, and what relative weights they should be given. There was universal agreement that modules from a standard student's first level (i.e. the first year of study in a three year full-time degree programme) were to be seen as preparatory rather than part of the main "honours programme", and so should not count at all for classification – thus only module grades achieved at the student's penultimate and final levels should be considered.[16] There was also general agreement, arising from massively increased practical experience of "strategic students" since modularisation (when detailed classification rules had first been published to students across the University), that *all of the modules taken in a student's final two levels* should normally count for classification – to do otherwise was simply to invite the strategic student to ignore those modules that would not count.[17] Finally, there was a clear consensus that modules taken at any particular level in a student's career should *in general* count equally, and hence that the determinant of a module's weight in the classification average should not be the level of the module itself, but

[16] Note that a module which is failed and then retaken at a subsequent level still counts as being of the initial level within that student's programme career. So a module taken in the first level and then resat at level 2 or 3 is not to be included in classification.

[17] Since modularisation, many departments had not included "elective" modules in classification – this gave rise to the notorious problem of "neglectives" (a term coined on a bulletin board set up by students to share cynical advice on the topic!), in which students would choose electives with a view to doing the minimum possible work to get the credits, since for their degree class, a bare pass would do just as well as a First Class grade. Even more serious problems arose in those departments which had adopted a policy of ignoring for classification purposes a student's *worst* results (e.g. the worst 20 or 40 credits) – here a student could with impunity choose to neglect *any* module at all (e.g. the ones they found most difficult) even if these modules happened to be central to the degree programme, and subsequent analysis of student performance made very clear that this had a very significant effect on student concentration of effort. By contrast with all this, those departments which (since modularisation) had included all upper-level modules in classification were unanimous about wishing to continue doing so. The overall lesson drawn from this experience across the University was that if it is an expectation that students should work seriously on 120 credits per year of study towards their degree, then 120 credits per year should count for classification of that degree. Hence in the agreed classification policy, the only upper-level modules not to count are those that are "supernumerary" – modules which a student chooses to take over and above the credit norm for the programme (e.g. an additional 10 credit IT elective, taken with future employment in mind rather than as part of the degree programme).

rather the level at which it was taken by the student concerned.[18] Thus the main issue of debate was the appropriate relative weighting of the penultimate and final levels within the classification calculation.

9. Weighting of Modules (A): Taking the Penultimate Level Seriously

Prior to modularisation, many degree programmes were designed as a closely integrated sequence of courses examined exclusively at the final level, and this naturally fostered an emphasis on the degree class as a measure of the student's "exit performance".[19] Even where courses were examined at the penultimate level, such examinations were often seen as merely preparatory or motivational, a recognition of the obvious fact that without such examinations, many students would be insufficiently motivated to treat the penultimate level with appropriate seriousness given its role as a foundation for the all-important final level. To enhance this motivation, penultimate level examinations were often counted into the degree classification calculation, but with only a low weighting (so that for example the penultimate level as a whole might count for just 20% or 25% of the overall result).[20]

Modularisation implied a serious reconsideration of this position, since now all teaching and assessment was to be packaged into semi-independent "modules", each with its own credit weighting which (in theory at least) was in direct proportion both to the hours of student time it demanded, and to the quantity of assessed work. In this new context, not only was it impossible to defer the explicit assessment of penultimate level material until the final level or to examine it implicitly through the student's final-level performance, but also, there was a strong *prima facie* case for giving the penultimate level assessments as a whole far greater weight than they had typically hitherto been given. Degree programmes were no longer to be seen as closely integrated wholes, but as structures formed from relatively self-contained "modules" with all assessment explicitly included within themselves and strictly proportional to the teaching and studying load which they entailed.[21]

---

[18] Several arguments can be given to back up this consensus. First, it is up to the degree programme rules to ensure an appropriate combination of modules at each level of study, so such appropriateness can reasonably be presupposed when it comes to classification. Secondly, what is appropriate can often depend on the programme (e.g. a Joint Honours student might well take at level 3 a module which Single Honours students take at level 2), so relying on the level of the module to determine the weighting will bring serious distortions. Thirdly, there is reason to view a student's performance over each level to some extent holistically – as the student matures intellectually, this will tend to be reflected in all the modules taken at that stage, so if the weighting is intended to reflect this sort of consideration, it should be applied accordingly across the board. Finally, weighting modules according to the level of the module would bring serious practical complications of calculation and comparability between students – for example, it would imply that the "denominator" of the classification average could vary significantly, and so how much each and every module counts for classification could turn out to depend on the precise combination of modules taken by the student.

[19] The term "exit velocity", often used in this context, is infelicitous because velocity is a *rate of change*. A student whose work rapidly improves during the final year (possibly from a very poor baseline) presumably has a higher "exit velocity" than a student whose work has been consistently excellent, but clearly does not merit a better result on that account.

[20] Analysis of past data confirms the importance for student motivation and performance of weighting penultimate level results significantly within degree classification. Students who take advantage of the irrelevance of penultimate level results by temporarily "slacking off" not only perform much worse than others at the penultimate level, but also perform worse at the final level – *even if their own performance dramatically improves between the two levels.* Thus even the student's *final level* performance benefits from a strong emphasis on the penultimate level.

[21] This does not imply, of course, that modular degree programmes cannot be seen as "integrated" in the sense of "well-structured". "Modularity" itself can be understood in various senses, but the one applicable here should be that of software engineering (where a well-designed computer program consists of modules that are "loosely coupled" but nevertheless

Hence where a student performed well in the 120 credits constituting the penultimate level, it now seemed seriously unjust that this performance should count for so little in the degree classification procedure, when it represented exactly the same amount of the student's effort and assessment as the 120 credits of the final level.

Another consideration which told in favour of giving greater weight to the student's penultimate level work derived from the changing nature of the students and their likely employment destinations. With a massive increase in the proportion of the population coming to university, and a relative decrease in the attractiveness of university employment, it became far less realistic to see the primary purpose of the degree as being to provide a foundation for academic research. Moreover the tendency of non-university employers, even in technical disciplines, to see the degree as a general education prior to detailed in-house training (rather than as a direct preparation for industrial work) again made it appropriate to reconsider the traditional emphasis on assessing the degree so predominantly by reference to the most research-oriented elements of the programme (typically the most advanced optional courses, and especially the final year project or dissertation). It is understandable why this traditional emphasis should have arisen, and also why it should remain attractive to research-oriented academics who may regret the change in focus. But though admittedly driven largely by non-academic considerations, it is hard to argue that an increased emphasis on the more general penultimate level is in fact detrimental to academic standards. First, the vast majority of students (both in discursive and technical disciplines, but especially the latter) perform significantly better at the final level, where they are able to concentrate on what interests them and can often discard areas of the subject that they find hard – hence increasing the weighting of the penultimate level makes the achievement of a good result if anything harder rather than easier. Secondly, it is precisely the specialist and research-oriented courses that typically carry most risk of biased or distorted assessment, owing to the small numbers of students on any such course, the wide variety of staff involved in teaching and supervision, and the relatively random distribution of students to supervisors. Statistical evidence clearly indicates that students' results on specialist options and projects or dissertations can be significantly dependent on the members of staff concerned,[22] so fairness and quality enhancement if anything suggest an increased emphasis, within degree classification, on those more general modules that are typically taken by large numbers of students at the penultimate level.

A final reason for weighting the penultimate level comparably with the final level in degree classification, which though less universal than those discussed above seems to have been quite influential in those departments to which it does apply, is that within a high proportion of degree programmes, many of the same modules can be taken at either level. In the case where a student does well at the penultimate level on a module which other students are taking at the final level, it seems extremely unjust if he or she is given significantly less credit than they are, purely on the basis of when the module was taken. Moreover in the context of a publicly

---

designed to interact logically and coherently) rather than that of furniture construction (where modularity can be simply a matter of size and shape, enabling all sorts of combinations to be fitted together without regard to functional coherence).

[22] This point applies particularly strongly in the case of final year projects or dissertations, which are inevitably more subject to these sorts of risk than other modules, and also typically carry sufficient weight (e.g. one third or even one half of the final level) to have a massive impact on degree classification, especially if the penultimate level counts for little. This is not in any way to deny that appropriate quality maintenance procedures can reduce these risks to acceptable levels, nor to question the educational value of including such major pieces of work within the degree programme. But such educational value by itself should not necessarily imply a predominant role in degree classification, especially once the risks have been recognised.

announced degree classification system, such a policy could have a seriously anomalous impact on student choice, motivating students quite artificially to defer until the final year those modules where their chief interests lie and on which they expect to perform best.

10. Weighting of Modules (B): Should the Final Level Nevertheless Predominate?

We have seen that in the context of a modern modular degree programme, there are several powerful reasons for treating a student's penultimate level as of comparable weight to the final level in classification. But let us now turn to the arguments that have been advanced for the opposite view.

One quite popular but specious argument is that since the final level builds on the penultimate level, it should therefore count more in classification. As it stands, this inference is simply a *non-sequitur* – the supposed link from *"A builds on B"* to *"A should count more than B"* is by no means evident, and needs to be argued for if it is to deserve serious consideration in the face of the points made above. Moreover as a general claim it is surely absurd, implying for example that competence in specific applications should always count for more than competence in the general theory on which those applications are based (or, at a more basic level, that in assessing a primary schoolchild's performance, competence in writing poetry should be given more weight than competence in writing *tout court*). In this sort of instance, as in many others, it seems on the contrary that more weight should be given to the fundamental core of the discipline than to the superstructure which builds on it.

A related argument, which again can seem initially quite appealing, is that the final level should count more in classification because it is intellectually more demanding than the penultimate level. However one might reasonably object, in the light of the previous section, that the premise of this argument is at least questionable as a general claim. The final level may indeed demand more intellectual *depth* than the penultimate level, but if so, since under a modular system it is not permitted to demand more overall student effort (given the standard norms on time required per credit), it will typically compensate for this increased depth with a reduction in *breadth*. Such evidence as is available does not seem to support the claim that this change of emphasis makes the final level overall more demanding for most students. But leaving this objection completely aside, why should it anyway be presumed that the most intellectually demanding assessments should automatically be given the greatest weight in classification? Of course a student who does well in them should, *other things being equal*, be given more credit than a student who does relatively poorly, but this applies just as much to the more elementary foundational assessments, and it is not clear why the latter should be held to count significantly less where the two criteria point in opposite directions. A student who does well in some difficult and specialised advanced options, but poorly on the fundamentals of the discipline, is not *self-evidently* more meritorious than a student who does correspondingly well on the foundational material but poorly on the advanced options.

Even where one student does better than another on *both* the foundational *and* the more advanced aspects of an academic programme, it does not follow that the latter should predominate in assessing the difference between them. This point is very widely recognised when setting and marking technical examinations: typically

a question might involve, say, four parts of increasing difficulty, and yet examiners consistently give just as many marks for the "easy" parts as for the "hard" parts, since only thus can a reasonable spread of results, properly indicative of the students' relative abilities, be achieved. Naturally this means that some marks are easier to obtain than others, but in a progressive discipline such variation is to be welcomed rather than regretted.[23] If examiners were to weight question parts within such examinations in ways that consistently reflected the effort they demand of *good* students (i.e. with very few marks for the "easy" parts, and lots of marks for the "hard" parts), then this would very obviously discriminate against the weaker students who can aspire only to master the easier material, and would be likely to produce a highly skewed (and probably bimodal) pattern of results seriously out of line with the examiners' judgement of the students. Such considerations commonly, and quite properly, lead examiners to give as many marks to the "easy" parts of a technical examination as to the "hard" parts which build on them, *even when the latter demand far more time and effort*. Surely, then, analogous considerations must be at least as applicable to the weighting of levels in modular degree classification, *where the penultimate and final levels are specifically designed to consume equal time and effort*.

11. Weighting of Modules (C): Exit Measurement, Balancing Incentives, and Special Skills

It might now seem that there can be no tenable reason for according the final level of a modular degree any more weight in classification than the penultimate level, given that both involve the same amount of student effort and assessment. However there remains a powerful argument on the other side, which becomes apparent if we turn our attention from the academic content of the modules themselves, and instead take account of the place of degree classification within the temporal context of the student's academic career. The point here is that the aim of university education is to *improve* the student's academic performance, and accordingly a major objective of degree classification is to reflect the level of performance that is *finally* achieved. Hence at least where a student's work does indeed improve, there is after all good reason for giving special weight to results achieved at the final level.[24] Thus the traditional idea of the degree class as a measure of "exit performance" still retains significant force.

It is important to note, however, that this argument has far less force in the relatively unusual situation where a student's performance *declines* in the final level, for in this case the discounting of the penultimate level would seem unjust in the light of the strong considerations discussed in the previous two sections. It is one thing to discount the student's penultimate level performance when the final level manifests a significant improvement; it is quite another to discount it when it represents the better half of the student's upper-level performance. Justice requires that a good performance at the penultimate level should be given due weight within classification, but justice is entirely consistent with more charitable treatment in cases where discounting of the penultimate level is in the student's interest.

---

[23] Since the "easy" parts are clearly identifiable, and typically pave the way for the "hard" parts, this is not at all like the unfair situation in which students are unknowingly given a choice between equally weighted questions of variable difficulty.

A similar conclusion is suggested by another argument which likewise appeals to the temporal aspect of the student's career, but this time focusing on motivation. Weighting the penultimate level equally with the final level provides the best possible incentive for taking the penultimate level seriously, but clearly such a fixed weighting could be seriously detrimental later on if the student performs badly. Suppose, for example, that he or she obtains only a middle Third Class result at the penultimate level – if this is destined to count equally with the final level, no matter how well the student performs there, then it makes the achievement of a 2:1 virtually impossible and might reasonably prompt him or her to aim no higher in the final level than whatever is needed to secure a modest 2:2 overall. However a policy which holds out the promise of greater weighting – say double weighting – on the final level subject to improved performance can remove this problem, enabling the student to aspire to a 2:1 overall by turning in a high 2:1 result at the final level. There is obviously a balance to be struck here, because the potential extra weighting of the final level must be enough to provide a strong incentive at that stage, but without undermining the incentive provided earlier at the penultimate level (e.g. a policy which *completely* discounted the penultimate level in cases of improvement would inevitably encourage students to neglect their penultimate level studies – hard work in a year's time will always be more attractive than hard work now, and the serious educational impact of this obvious psychological point is emphasised by the analysis of past data mentioned in footnote 20). After consultation, a very clear consensus emerged that *double* weighting of the final level in cases of improvement was indeed the appropriate balance.

Double weighting of the final level pointed up, however, the crucial importance of ensuring that modules taken at that stage should be of an appropriate academic difficulty – the practice of allowing some finalists an unrestricted choice of "elective" modules could not be maintained if these modules were to carry such weight within classification. The obvious solution was to debar third (and fourth) level students entirely from taking any level-1 modules, but this would have had the clearly undesirable effect of preventing them from taking a range of life-skills and career-oriented modules from which they particularly stood to benefit. Students about to leave university for the job market are well-advised to acquire information technology, language, and numeracy skills if they do not already have them, and to attend special modules provided by the University Careers Service to prepare themselves for employment. Such modules are unlikely to be particularly demanding in terms of intellectual *depth* (and therefore may be just as suitable for level-1 students as for finalists), but this should not prevent them from being taken at the final level. After consultation with all the faculties, an elegant compromise to this tension was reached, whereby those modules teaching life and career skills of recognised importance to final-level students should be made available to them as "Special Skills Electives". A list of modules eligible for this status would be maintained, and these would be the only level-1 modules allowed to students beyond level 2. Even when taken in the final level, however, these modules would *not* be subject to potential double-weighting within the classification calculation – for classification purposes, they would count in exactly the same way as if they had been taken at the penultimate level.

---

[24] It is significant that in those departments which since modularisation had adopted a strict policy of equal weighting of levels (principally in the Arts and Social Sciences), this failure to "reward" final level improvement was the only serious concern raised about that policy.

Weighting of Levels and the Classification Average

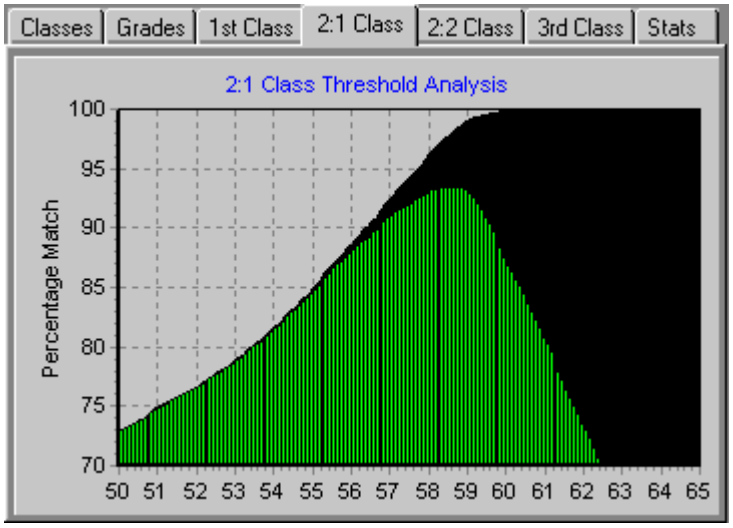
It was agreed that for classification purposes, the student's average performance should be calculated *both* on the basis of a 1 to 1 weighting of penultimate to final level, *and* on the basis of a 1 to 2 weighting. Then the student should be classified according to whichever of these overall averages is the better. (This is equivalent to treating both levels equally if the student does worse in the final level, but giving the final level double weight if the student improves.) The better of the two averages is accordingly called the student's *Classification Average.*

All modules taken at the penultimate and final levels should be included in these averages, with the sole exception of "supernumerary" modules (i.e. those taken over and above the normal credits for the program). To avoid inappropriate double weighting, level-1 modules are in general forbidden beyond the student's second level. However certain level-1 modules that provide recognised life and career skills may be taken at the final level, in which case they are called "Special Skills Electives" and count for classification as though taken at the penultimate level (i.e. they cannot be double-weighted).

12. Fixing the Formal Classification Thresholds

With the Module Grade Scale and the appropriate weighting of module grades having both been agreed, it remained to decided what thresholds should be applied to each student's Classification Average (i.e. appropriately weighted grade average) in determining the appropriate class of degree. Here the benchmark was to be conformity with the University's previous classification outcomes, since it was universally agreed that the new system should yield results broadly comparable with those of past classification standards. Accordingly, a special computer program was written to investigate these past standards in a systematic manner, enabling all post-modular classifications to be collated and analysed collectively in great detail. This analysis was also compared with the output of another computer program, written to analyse the University's *pre-modular* classification outcomes from the academic sessions 1992-93 and 1993-94 (the only two sessions for which relatively reliable data were available from the current Student Information System). Both analyses yielded similar results, thus providing assurance that consistency over time was being maintained.

To illustrate the method of deriving appropriate classification thresholds, here is one of the graphs generated by the post-modular analysis program, based on degree classification outcomes over the entire University in the sessions 1995-98, and involving in all over 16,000 students. It shows the results of "2:1 Class Threshold Analysis", in which the classifications that would result from using the newly agreed methods of module weighting (as discussed in sections 8 to 11 above), together with a range of possible 2:1 thresholds, are compared against the classifications actually awarded:



This indicates that the best *automatic* match with actual practice can be achieved by choosing a formal 2:1 threshold of between 58·1 and 58·9 – if classification were to be done *purely* by reference to whether each student's Classification Average reaches the threshold, with absolutely no application of positive examiners' discretion, then any threshold in this range would achieve a match of around 93·4% (as shown by the lower, lighter coloured graph). The remaining 6·6% of students, for whom such automatic classification would not match with their actual result, fall into two categories. First, there are those who were in fact given a 2:1 but whose Classification Average *did not* reach the threshold in question – these are indicated by the black area. Secondly, there are those who were in fact given a 2:2 but whose Classification Average *did* reach the threshold – these are indicated by the gap above the black area. As the threshold gets more demanding (i.e. moving from left to right across the graph), there are of course more 2:1 students who fail to reach the threshold, and fewer 2:2 students who succeed in doing so.

If we now take positive examiners' discretion into account, however, we can see that the extent of *automatic* matching with past practice is not the whole story. For a student who fails to reach the formal 2:1 threshold does not therefore *have to be* denied a 2:1 degree – indeed the main point of holding classification meetings is precisely to discuss factors which might justify awarding the higher class to students whose results are insufficient to qualify "as of right" (the historical application of such discretion presumably accounts for a high proportion of the students in the black range above the peak of the lower, lighter coloured, area – students whose results were not sufficient for an automatic 2:1, but who were judged worthy of a 2:1 nonetheless). The optimum choice of threshold, therefore, is not the one which gives the best *automatic* match with past practice, but rather, the one which is likely to offer the best *overall* match after positive examiners' discretion has played its part, and without requiring excessive application of such discretion. In terms of the graph, this means that we should seek a (preferably round-numbered) value at which the lower (lighter coloured) area is very near to its maximum, and the top of the black range very near to 100% – from examination of the graph, a threshold value

of 59·0 seems optimal.[25] Similar analysis on the First Class boundary yields an optimal threshold value of 68·5, and corresponding choice of lower thresholds completes the following sequence, neatly spaced at even intervals of 9·5:

The Formal Classification Thresholds

The following formal thresholds were agreed for classification, meaning that any student whose Classification Average (i.e. weighted module grade average) reaches a particular threshold should be awarded at least the corresponding class of degree:

First Class: 68·5 Upper Second: 59·0 Lower Second: 49·5 Third Class: 40·0

The numbers here might occasion some surprise for those expecting the thresholds to lie exactly 1·0 apart, but in fact this makes excellent sense theoretically as well as practically. The point is that the values which are to be measured against these thresholds are *averages* of many module grades, typically between 12 and 24 of them. And statistically, averages have a strong "central tendency": it is far more unusual to find an average which is distant from the overall population mean than an individual which is similarly distant (for example it is far more unusual to find a football team with an average height over 2 metres than it is to find a single footballer over 2 metres). So while an individual module grade of 70 might be moderately impressive, an overall *grade average* of 70 is very much more impressive – around 13% of module grades achieve the former, while only 5% of students achieve the latter. Likewise at the other end of the scale, an individual module grade of 40 is certainly poor (in the bottom 7% or so of module grades), but an overall grade average of 40 is far worse (being "achieved" by only around 0·6% of students). This statistical point, whilst not usually recognised explicitly, is in fact implicitly very familiar to most examiners, as shown by the following illustration. It would be almost universally agreed that a student who scores 70 on all 24 modules thereby achieves not only a First Class degree, but a *very fine* First; whereas a student who scores 40 on all 24 modules thereby achieves a *very poor* Third. The implication of this illustration for any mark-averaging classification system is clear, though not widely recognised: whereas a module grade of 40 should only just scrape above the threshold for a Third Class Classification Average, a module grade of 70 should be comfortably above the minimum First Class Classification Average. It is just this that the sequence of thresholds shown above achieves, producing an elegant conformity between the empirical analysis of past classification practice and the appropriate theoretical statistical constraints.[26]

[25] For a far more detailed analysis of how the value of 59·0 turns out to be optimal, see the appendix section entitled "Computerised Comparison of Possible Classification Thresholds as Applied to Past Classification Outcomes".

[26] The issues sketched in this paragraph are discussed in more detail, with further supporting data, in the appendix section entitled "Practical and Statistical Considerations on the Spacing of Classification Thresholds".

13. Examiners' Discretion, Vivas, and the Adjustment of Module Grades

A student whose Classification Average reaches a formal classification threshold can expect to be awarded at least the corresponding class of degree, as long as he or she has not violated the published provisions of the degree programme in question (for example by failing to pass, or to reach an adequate grade in, enough credits or a specifically required module). But, as noted in the previous section, it does not follow that a student who narrowly fails to achieve that threshold will usually be denied that class. For it has long been standard practice across most university departments to consider individually those students whose Classification Average falls closely below any formal threshold, with a view to possibly awarding them the higher class despite their failure to qualify for it "as of right". Such favourable discretion can take either of two forms:

(a) "Normal" discretion, where a student's Classification Average is considered sufficiently close to the threshold for him or her to be awarded the higher class of degree on the basis of considerations such as "class profile", concentration of grades, steady improvement, performance in central modules, etc.

(b) "Special" discretion, where there is evidence that some grades on which the Classification Average is based do not provide a faithful representation of the student's general standard of performance. Serious medical or personal problems, or a radical change of attitude to study reflected by an exceptional improvement in final year performance, are the most commonly invoked grounds here.

It has long been standard for such discretionary processes to be treated semi-formally, with "normal" discretion being contemplated only for students whose Classification Average lay within some precise range below each formal threshold, and "special" discretion being permitted only on particular approved grounds (e.g. written medical evidence). But computer analysis of past results indicated that within the University of Leeds there had nevertheless been wide variation in practice – some departments had apparently applied positive discretion to the great majority of students falling within 1 or 2 marks of the formal classification thresholds, while others (even if notionally following *identical* published rules) had interpreted the formal thresholds far more strictly. Clearly this variation was potentially very unfair to students, so a cross-University consensus had to be reached.

Fortunately the newly agreed formal classification thresholds made agreement on the appropriate margins for "normal" discretion far more straightforward. For those departments which had in the past contemplated such discretion over a wide range (e.g. up to 2 whole marks below the formal thresholds) had typically done so as a result of using the traditional "round number" thresholds that at the First Class and 2:1 boundaries were far too high. Thus departments applying a formal threshold of 70 for a First might be considering students for positive discretion if their Classification Average fell up to 2 marks below this (i.e. as low as 68·0); while departments applying a formal threshold of 68·5 would typically allow discretion only within a far narrower range, perhaps half a mark (i.e. again as low as 68·0). The upshot was that these various departments, whose rules and thresholds seemed superficially very different, were often in practice classifying to very similar standards. Agreement to the formal classification thresholds given in the previous section thus opened the way to agreement on a narrow margin for "normal" discretion, standardly 0·5 of a mark (though extendable to 1·0 in cross-departmental degree programmes where there is relatively limited scope for vivas and re-reading of scripts). This meant that to be eligible for a First on the basis of "normal" discretion, a student must henceforth

achieve a Classification Average of 68·0; for a 2:1, 58·5; and for a 2:2, 49·0. None of this, however, ruled out the more generous operation of "special" discretion in cases where documented medical evidence or suchlike could be brought to bear.

Both "normal" and "special" discretion are to be distinguished sharply from a third form of discretion, in which boards of examiners consider the use of *viva voce* examinations to provide "sub-borderline" students with an opportunity to improve their grades. *Viva voce* examination, just like any other form of additional assessment (e.g. re-reading of scripts by external examiners), naturally implies the possibility of grade revision and hence in some cases a change to the overall class awarded. But the "normal" and "special" forms of discretion discussed above are conceptually quite distinct from this: unlike vivas they involve no new assessment and hence can justify no change of module grades; they simply involve the award of a degree class higher than the one which the student's grades would have achieved solely by reference to the formal class thresholds.

Many departments traditionally use *viva voce* examinations only in the most exceptional cases, while some others, at the opposite extreme, use them as a standard form of assessment for all students. Departments falling between these two extremes, which accord vivas a major role but only in resolving the classification of sub-borderline students, are likely to be led by considerations of quality assurance and fairness to define internally agreed "viva thresholds", so that any student who achieves a Classification Average within some specified range (and not just those who for some reason catch the examiners' attention) will be given the chance of a viva to improve his or her grades. Thus in principle a department of this sort might employ three different thresholds close to the First/2:1 borderline, say 68·5 (the "formal threshold" published to students, and meriting a First Class award as of right), 68·0 (the "discretionary threshold" which determines the lowest Classification Average consistent with discretionary award of a First) and, say, 67·0 (the "viva threshold", such that any student who achieves this average will at least be viva'd with a view to possibly raising the average and achieving a First). Although there is an obvious quality requirement for broad consistency of formal and discretionary thresholds across departments and faculties, the same is not true of viva thresholds, as these introduce an additional form of assessment which may be more or less appropriate (or indeed practicable) within different disciplines. Hence nothing in the new degree regulations prevents departments and faculties agreeing their own norms on the operation of *viva voce* examinations.[27]

[27] However it is important to stress that the practice (anecdotally quite common in the past) of seeing the viva as primarily an opportunity to assess "the quality of the student's mind", is now generally agreed to be totally unacceptable. The viva must provide assessment relating to the modules taken by the student, and any adjustment to those module grades must be firmly based on what the viva reveals about the student's *relevant* knowledge and understanding. Conversely, if the viva reveals no such *relevant* information, and hence justifies no change in the module grades, then it cannot provide a basis for changing the student's degree class, even if the examiners derive the impression of a student with a "First Class mind". The degree class is intended to provide an objective measure of the quality of the student's performance on the modules that constitute the programme; it is not, therefore, appropriately used to record the examiners' subjective opinion of the student's intellect except in so far as the quality of that intellect has been manifested within the content of the programme.

Normal Discretion, Special Discretion, Vivas, and Reassessment

It was agreed that in general, any student whose Classification Average falls within a range 0·5 marks below a formal classification threshold should be considered for "normal" discretion, and no student outside this range should be so considered. "Special" discretion remains applicable in any case where appropriately documented medical or other evidence is available, and reassessment by *viva voce* examination or by re-reading of scripts (in each case potentially leading to a change of module grade) remains a matter for the department or faculty concerned.

14. The Introduction of Decimal Grades

The only prominent feature of the new classification system that remains to be explained is its use of "decimal grades" – that is, the expression of "Classification Grades" and "Classification Averages" using a scale that runs from 2·0 to 9·0 instead of the module grade scale running from 20 to 90. This feature is superficially quite prominent, but since it has little impact on the system's fundamental logic and principles, and would have introduced some complexity into the discussion of those principles, it has been left until last.

Although they usually go together, the numerical value of a module grade and its contribution to the Classification Average are conceptually distinguishable. This is most clear in the case of the module grades "ABS" (absent) and "NSA" (no serious attempt), neither of which has any numerical value as it stands, but which count for classification as though they were module grades of "0" and "10" respectively. (However this last way of putting the matter is potentially confusing, because on the new module grade scale there is no such grade as "10", and the module grade "NSA" has a specially defined criterion of application which is intended to distinguish it quite clearly from any numerical grade.[28]) Even with other module grades, however, the distinction between the grade's numerical value and its contribution to classification is significant because the two relate to different scales, the module grade scale and the scale of classification thresholds respectively.

A second reason for introducing a distinct scale to represent the contribution that a module makes towards classification is the continuing desire of some departments to mark on the traditional 0-100 scale. In those departments where marks above 90 have traditionally been fairly common (such as Physics), there is naturally a worry that a move to the new 20-90 module grade scale might disadvantage Leeds students in competition for grants with students from elsewhere. Hence these departments are continuing to mark on the 0-100 scale at least for the present, and although a translation table has been published to ensure that the two scales can

[28] As explained in Section 6 above, "NSA" was introduced as an explicitly *punitive* grade, intended to target students who neglect modules rather than those who try hard but cannot cope. Hence it was agreed that the relevant criterion should be based not on the *quality* of work done, but on the *quantity* of work genuinely attempted. NSA is to be applied in cases where a student's performance on a module, while including *some* work worthy of assessment (making "ABS" or "0" inappropriate), nevertheless includes so little such work that it would have been insufficient to pass the module even if the student had scored maximum marks on the parts that they genuinely attempted.

properly be related to each other, this clearly brings some potential for ambiguity over the 20-29 and 81-90 ranges where the two scales do not go in step.[29]

For these reasons, the concept of a "Classification Grade" was introduced, this being a numerical representation of the contribution of a module to the Classification Average. To ensure that these Classification Grades remain straightforwardly distinguishable from module grades, it was agreed that they should be expressed – always with one digit before and one digit after the decimal point – on a scale whose main body runs from 2·0 to 9·0 (exactly matching the main body of the module grade scale from 20 to 90), and with two additional points of 0·0 and 1·0 below this main scale to provide appropriate Classification Grades for "ABS" and "NSA" respectively. Classification Averages are to be calculated straightforwardly from these decimal values, but then expressed with an additional digit after the decimal point to reflect the additional accuracy applicable to such an average. Accordingly, the classification thresholds were redefined in the obvious way, becoming 6·85 for First Class, 5·90 for 2:1, 4·95 for 2:2 and 4·00 for Third Class.

Although initially the introduction of decimal grades might seem to make the new classification system more complex and confusing, its real impact is quite the reverse. Decimal grades enforce a clear visible distinction between three different kinds of numbers which can be used within classification:

Module grades	e.g. 37, 53, 84	always expressed as integers
Classification Grades	e.g. 3·7, 5·3, 8·4	always expressed to one decimal place
Classification Averages	e.g. 4·00, 5·32, 6·80	always expressed to two decimal places

Quite apart from the conceptual clarity which this distinction brings, it also greatly reduces the chances of undetected data errors arising in a context of increased automation and computerised processing of degree classification calculations.

Module Grades and Classification Grades

To enforce a systematic distinction between module grades and the Classification Grades, it was decided that the latter should be expressed as decimal numbers on a scale between 2·0 and 9·0, corresponding point for point with the standard module grade scale between 20 and 90. The classification thresholds are also changed accordingly (from 68·5 etc, to 6·85 etc), so that all degree class calculation is to take place on a scale which is quite distinct from the module grade scale. This "decimal grade" provision completes the definition of this system of degree classification.

[29] The two scales correspond exactly over the 30-80 range, with linear interpolation applying outside that range. The grade of "NSA" is quite independent of these scales, and is therefore to be sharply distinguished from a grade of "10" on the 0-100 scale (emphasising further the point made in the previous paragraph). In future, *any* non-zero numeric module grade is to be taken to imply a "serious attempt" – hence a non-serious attempt must be returned explicitly as "NSA" rather than as any number.