Working Paper Series
ISSN 1170-487X

A non-linear, criterion-referenced grading scheme for a computer literacy course.

by Bill Rogers, Phillip Treweek, Sally Jo Cunningham

Working Paper 96/7

April 1996

Department of Computer Science
The University of Waikato
Private Bag 3105

A non-linear, criterion-referenced grading scheme for a computer literacy course

Bill Rogers, Phillip Treweek, Sally Jo Cunningham
Department of Computer Science
University of Waikato
Hamilton, New Zealand
{coms0108, coms1165, sallyjo} @ cs.waikato.ac.nz

Abstract: When we revamped our first year computer literacy course, we moved from one major grading paradigm (norm-based) to another (criterion-referenced). We also introduced a novel sliding curve formula for calculating grades that we hoped would encourage more students to finish the course, and to complete all of the coursework. This paper explores the effectiveness of that new grading scheme.

1. Introduction

From 1989 to 1994 our computer science department offered a computer literacy course, called "Computing Applications", that followed a pattern common throughout the world: we concentrated on the "big three" applications (word processors, spreadsheets, and databases). Students completed practical assignments with these programs, attended lectures explaining their use and heard a bit about contextual topics such as computing ethics and business uses for computers. The course was aimed primarily at non-majors, but was also taken by majors. The final course grade was based on a weighted averaging of applied computer work and tests (including a formal final examination).

As the years passed, we became more and more dissatisfied with this course model. It was difficult to reconcile the often conflicting needs of majors and non-majors in a single applications class. In addition, an increasing percentage of our incoming students had already learned one or more of the big three applications in a high school computing course, and were not being challenged by our introductory material. At the same time, we still felt a need to cater for our novice students, who had never worked with a computer before. Finally, we wanted to tailor the course to the needs of different majors, to allow students to apply different types of software to the problems of their own discipline of interest. We needed a greater degree of flexibility in both the topics and the relative difficulty of assignments to better serve the mix of students in our course.

In 1995 we trialed a radical redesign of Computer Applications. The new course, named The Computing Experience, was based entirely on practical work; students selected a set of software from a wide range of options and worked through tutorials on the packages. We limited the course to non-majors, with the following stated goals: to offer students practical skills in using computers; to encourage an appreciation of the general capabilities of computers and an understanding of how computers could be used to support work in their particular major; and — perhaps most importantly — to provide an opportunity to have fun with computers and see the enjoyable, challenging side of computing. To maintain a course focus on *using* computers, rather than learning lecture material, we deliberately eliminated the midterm test and final exam.

As a consequence, we moved the course from a grading scheme whose most heavily weighted assessment item, the final exam, was norm-based (in which marks are intended to assess the differences between the achievements of students) to a scheme that was primarily criterion-referenced (in which marks assess the difference between a given student's achievement and a list of learning outcomes). We also introduced a novel non-linear formula for calculating course grades. This paper explores the effects of the new grading scheme on the distribution of student marks, student performance in the course, and political repercussions in the university.

The paper is organized as follows: Section 2 provides a literature review of common and unusual grading schemes; Section 3 describes the grading schemes employed for Computer Applications and The Computing Experience; Section 4 presents typical student results and responses for the grading schemes; and Section 5 discusses the outcomes of our experiment with the new marking criteria.

2. Literature review

The current system of awarding letter grades or expressing student accomplishment in a course as a number scored out of 100 is relatively new educationally, coming into common usage in the nineteenth century [Durm, 1993]. The numeric scale and methods for calculating student grades have received considerable attention over the years, perhaps in part because of the association of numbers and formulae with scientific, objective precision. Two grading techniques, norm-based and criterion-referenced, have emerged as the major marking paradigms, with norm-based being the most widely accepted. In addition, non-cognitive student qualities are often tacitly included in the calculation of both norm-based and criterion-referenced grades. These general grading techniques are discussed below, followed by an overview of common and uncommon formulae reported in the literature for implementing these marking paradigms.

Norm-based grading

In a norm-based system, students are graded by comparison to each other. The expectation is that in a sufficiently large student population, the distribution of student grades will form a normal or "bell" curve. This distribution is usually created in one of two ways:

- Grades are directly calculated by reference to standard deviations from the mean; marks within a single standard deviation receive a "C", marks higher than one standard deviation but less than two receive a "B", and so forth.

- The assessment method and tests are "corrected" to produce a normal curve. For example, exam questions that "too many" students answer correctly are deleted, or the number of marks awarded to given pieces of work are manipulated until the expected proportion of students receive each grade.

Norm-based grading is predicated on the assumption that for a given task or course, only a few can do well, most people will perform adequately, and some will fail. The unfortunate corollary is that it is expected that the majority will never learn to perform excellently ([Rowntree, 87]; [Wiggins, 88]); indeed, there is evidence that as the majority moves closer to the educational goals, the goals are moved so as to maintain the curve [Rowntree, 87]. The expectation of failure or at best mediocrity can become institutionalized in students. Moreover, this assessment method can actually depress grades if students can self-select whether they will complete a course or degree program. As students of lower ability drop out, the distribution remains the same; students whose work would formerly be classified as adequate are forced to fill out the bottom of the bell curve [Wood, 1994]. Note that an ideal question on a norm-based test is one that is failed by about half of the students. This type of question fulfils the main objective of a norm-based test: to rank students by relative ability or expertise. However, this type of test construction implies that the norm-based assessment is less able to provide insight into the specific skills attained by a given student (ie, whether or not the student has achieved a particular objective) [Rowntree, 1987].

On the positive side, norm-based assessment by its very nature allows students to view their progress in comparison with others, and to evaluate their work in the wide context of *relative* achievement. From a teacher's point of view, norm-based tests may be particularly appropriate when students with a broad range of abilities are expected to progress through material at roughly the same pace. In this case, the basic normative assumption — that some material will be too difficult for some students to learn *in a given time period* — holds true.

Criterion-referenced assessment

With this assessment technique, teachers codify a list of learning or performance criteria related to the teaching objectives. Students are then assessed according to how well they meet these criteria. This method differs from norm-based marking in that there is no a priori assumption about how many of the students will be able to meet the criteria; indeed, the goal is to have as many of the students as possible achieve high levels of mastery.

The problem of specifying and measuring against appropriate absolute scales is, of course, difficult. At its worst, criterion-based assessment can degenerate into a dreary "teaching to the test", or can imply that the curriculum is "dumbed down" to the lowest common denominator. At its best, this technique holds up a standard for excellence that can be achieved or approached by most students, with the understanding that some students will take longer than others.

Non-cognitive criteria

The general agreement among university teachers is that grades should be based solely on a student's achievements (whether this achievement is measured absolutely or in comparison with the work of others). In practise, however, teachers commonly incorporate a number of non-cognitive, and relatively intangible, criteria into their grading scheme: the amount of effort that the student has put into the course, the student's perceived background and general ability level, the positive or negative attitude of the student, etc. [Cross et al, 1993]. Usually these extra criteria are informally used to "adjust" borderline grades, and are rarely discussed (or even admitted!) to students. One semi-serious attempt to make explicit this covert practise has been reported in the literature: Hinely [1968] proposed a "handicapping formula" that increases a raw test score in inverse proportion to a student's IQ, current grade point average, and last test score. The first item in the formula, the IQ, allows the instructor to "grade up" with the less able students, and the latter two items incorporate the practise of rewarding improvement or punishing a decline in achievement.

If used carefully, this type of grading scheme can have a positive effect on students — by, for example, encouraging a "trier" with a higher mark than would otherwise be awarded, or shaming a "slacker" with a lower grade. Beyond the problem of justifying the inclusion of completely subjective criteria that are often tied to a teacher's personal reaction to a student, the use of criteria that are not achievement-based has other objectionable features. One is that students may be mislead as to their likely appraisal by other assessors who do not know the student personally; another is that the grade may be misinterpreted as a measure of achievement or competence that the student may not have attained [Rowntree, 1987].

Formulae for calculating grades

Most implementations of the above schemes award a final course grade based on an averaging of internal course marks. In many cases the actual formula employed is a linear weighted average, allowing some pieces of work to have a greater influence on the final grade than others. Other types of formulae have been proposed:

- *median, rather than mean, grading*. Although the mean of a student's marks is most commonly used to calculate that student's final grade in a course, another measure of central tendency may be more appropriate: the median. The imprecise nature of grading means that marks awarded are essentially ordinal — and statistically, a median is more appropriate a summary than the mean for ordinal values. Pedagogically, a number of advantages are claimed for median grading: teachers can set higher standards while still awarding an acceptable proportion of passing grades; students are not unduly penalized for a single poor performance, which also results in a lower anxiety level at test time; and grades tend to be higher than with mean-based calculations, which in turn motivates C students to work harder (in the expectation of achieving higher grades) ([Wright, 1989]; [Wright, 1994]).

- *sliding grades*. All students attempt the same type of work, but students are classified by native ability level. The actual formula used to calculate a final grade will vary by this level: for example, the least able may have a grade calculated by

  G = 1/3 absolute mastery + 1/3 effort + 1/3 progress

  Students of moderate ability would have grades calculated by

  G = 0.5 * absolute mastery + 0.25 effort + 0.25 progress

  and students with the greatest native ability would be judged by

  G = 0.8 * absolute mastery + 0.1 * effort + 0.1 * progress

  The analogy here is to sports. As we would not expect a flyweight boxer to go up against a heavyweight, so it is appropriate to judge differently-abled students by different criteria. Moreover, students should not be absolutely classified by ability level for all time, but should be occasionally "re-weighed" and encouraged to move to the more stringent grading schemes as warranted by an increase in their skills [Wiggins, 1988].

- *postmodern grading*. Grade inflation, or an increase in the average student grade, has spread to the extent that in the US the most frequently given mark is an "A" [Farley, 1995]. Rather than ascribing this trend to an erosion of standards, we might instead attribute it to a new type of mindset: the postmodern worldview. Under the postmodern educational paradigm, an "A" grade is

earned by students who "learn to learn", rather than being bestowed as a reward for performing against a teacher-constructed standard. Given this new definition of meritorious work, what has been perceived as grade inflation actually indicates that our teaching has become more effective:

"A large number of As, therefore, signals many students' demonstration of learning, implicitly suggesting strong and effective pedagogy. A high grade distribution indicates that the teacher and students have been able to structure classroom activities in such a way as to facilitate learning by may students and that many students are sufficiently self-directed to employ these opportunities to learn." [Bilimoria, 1995, p. 452]

Postmodern marking is difficult to summarize with a simple formula:

"Postmodern teaching and evaluation practices, in contrast, blur the distinction between knowledge generation and dissemination. The process of emergent cocreation of knowledge among participants in the learning endeavor suggests the relational construction of knowledge...Thus the modern educational practise follow[ed] the principle of learning first and testing later — the knowledge that has already been created is learned and then evaluated. This knowledge is largely static, historical, and fragmented by disciplinary bounds. Postmodern education follows the principle of testing first and learning later — experiences generate knowledge, which is described by personal learnings. The knowledge that is created is dynamic, emergent, and holistic." [Bilimoria, 1995, pp. 454-455].

- *fuzzy grades*. Since the object of measurement — human achievement or performance — cannot be readily reduced to a crisp definition, fuzzy logic should be employed instead. Fuzzy membership functions into the numeric grading range for A, B, C, etc. can be set in any way appropriate to the course or institution; Echauz and Vachtsevanos [1995] suggest using a poll of departmental teachers to provide institutionally-focussed values instantiating a fuzzy measure that incorporates individual student absolute and relative performance, teacher effectiveness, difficulty of material, time allowed for the completion of assessed material, expected range and distribution of grades, and validity of the testing/assessment measures. For some courses, this technique can also satisfy the *curriculum based* grading method described below.

- *curriculum based schemes*. A university professor in Louisiana (who shall otherwise remain unidentified) bases his grading formula on the type of mathematics covered in his course. Since he usually teaches advanced undergraduate or graduate courses, the formula employed can be quite esoteric indeed. The advantage he claims to this method is that students are highly motivated to learn the content of the course, so that they can understand how their final grade has been awarded!

3. Grading schemes for Computing Applications and The Computing Experience

In this section, we describe the grading schemes used in both computer literacy courses, and discuss the effects that each scheme had on the distribution of student grades.

Computer Applications

The grading scheme for our original Computer Applications course was primarily norm-based and calculated by a weighted average. Specifically, we used the following formula to calculate a final grade:

$$G = 0.28 * (\text{four assignment marks}) + 0.05 * \text{mid-term test} + 0.67 * \text{final exam}$$

The four assignments involved creating small projects with a word processor, spreadsheet, and database, and finally creating an integrated report using all three pieces of software. Each assignment was expected to take five hours to complete. In addition, the students were expected to complete a series of 18 one hour practicals, which were themselves ungraded but were useful in completing the assignments or were featured on the examination. It was a course requirement that at least 14 of the 18 practicals had to be completed.

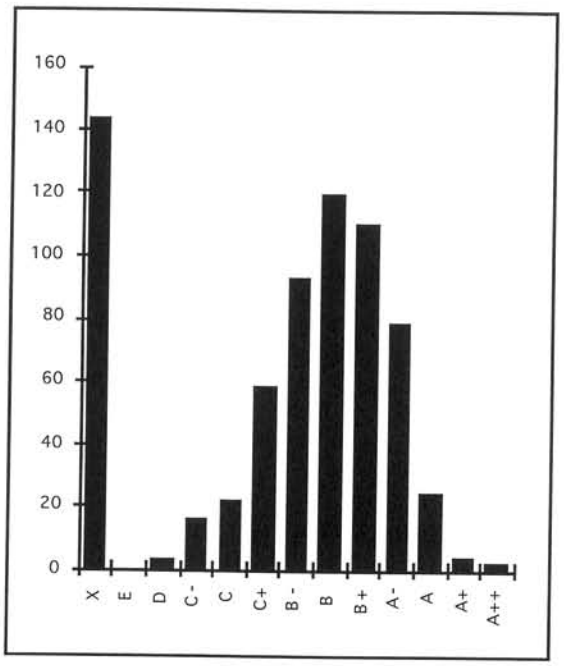

Figure 1. A final grade distribution for Computer Applications

Figure 1 presents a bar chart of the grade spread for the 1993 whole year version of Computing Applications. The grade distribution is typical for the course, and roughly conforms to a normal curve — as would be expected for norm-based assessment. Closer examination of the individual components of the course grades shows that most students did well on the computing assignments (Figure 2). These assignments were marked according to a criterion-referenced scheme, and the skewing of marks to the higher grades is typical of that type of assessment. The heavily weighted final examination, and to a lesser extent the mid-term test, are the main factors of the grading formula that spread the final marks into a bell curve.

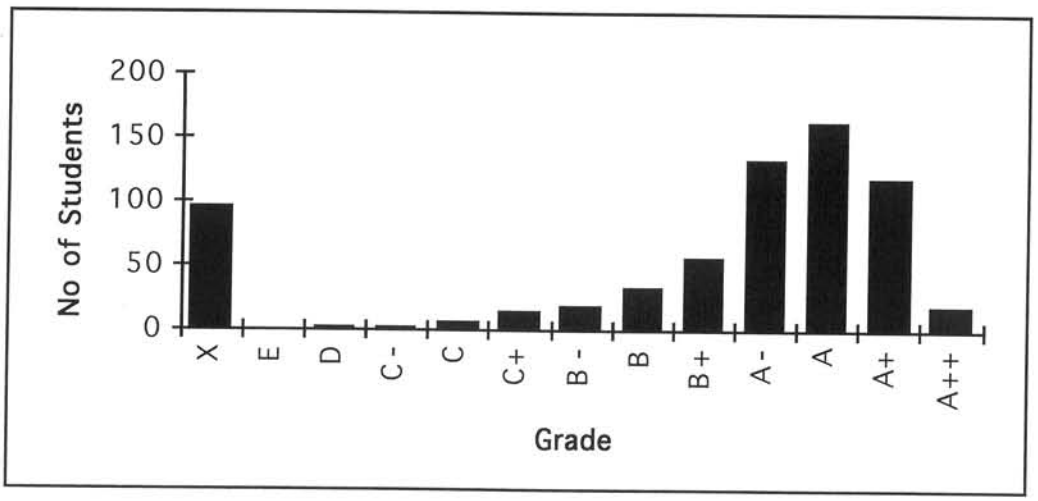

Figure 2. Assignment grade distribution for Computer Applications

A relatively high proportion (21.2%) of students were awarded an "X" grade (denoting that the student failed to complete the course, or informally dropped the class). That one fifth of our students failed to complete the course was of concern to us. Another worrying problem was the relatively high proportion of students who completed the course, but who did not complete all of the practicals (Figure 3). Apparently, our requirement that at least 14 of the 18 practical be completed was being interpreted by many students as giving permission to skip up to four of the practicals. In contrast, the fact that assignments were being graded, and counted for a reasonably large percentage of the final grade, seems to have encouraged students to complete all four of them (Figure 4). This result fits in with the general observation that students have been socialized to work primarily for grades, and to view

ungraded work as less important (for a particularly poignant discussion of this phenomenon, see [Kuby, 1994]).

number of practicals completed	<= 13	14	15	16	17	18
number of students	92	247	113	80	54	66
percentage of students	14.1	37.9	17.3	12.3	8.3	10.1

Figure 3. Distribution of student practical completion in Computer Applications

assignment	0	1	2	3	4
number of students	40	27	14	47	524
percentage of students	6.1	4.2	2.1	7.2	80.4

Figure 4. Distribution of student assignment completion in Computer Applications

The Computing Experience

The Computing Experience is based entirely on modules; lectures are given on general topics of interest and to provide context for novice computer users, but attendance at lectures is not required and the course does not have any tests or a final exam. Each module is described in a self-guided exercise, and is expected to require approximately five hours of hands-on work to complete. At the beginning of the semester students select a number of modules to complete, based on their personal educational needs and interests. These practical modules are of two types: P1, providing a guided introduction to a particular software package; and P2, offering the opportunity to complete a small project using software previously explored in a P1. P2 exercises are more demanding, and are therefore weighted more heavily when calculating a final grade. The normal expectation for the course is that a student will complete 9 practicals, with a mixture of P1s and P2s.

P1s are graded in the lab. Each practical is accompanied by a list of things the student should have done, or be able to do. When a student has worked through a practical they ask a demonstrator to *verify* their work. The demonstrator will work through some or all of the items on the list. They may look at finished work, or ask a student do give a practical demonstration of their ability to achieve some result. The demonstrators have considerable latitude in this testing process. If they have been working closely with a student, they may already know what the student can do, and the verification may be very quick. If the student has worked alone, more careful testing may be required, although this should never take more than ten minutes. The result is a grade in the range zero to four. Four is awarded for a demonstrated mastery of the material, and is overwhelmingly the most commonly awarded grade. Two's or three's can be given where substantial parts of the practical exercise have not been completed, but a demonstrator will more often advise a student to do a little more work and reattempt verification.

P1s usually involve a series of experiments. P2s involve work towards a finished product — either on paper, or on computer (eg: a Hypercard stack). The finished product is handed in, and marked by tutors against a specified list of criteria. Percentage grades are awarded. The distribution is spread over the ABC range, and failing grades are rare.

The formula used to derive the overall course score is the novel aspect. We were particularly concerned that students be strongly encouraged to complete 9 practicals. A linear grading scheme which might offer a pass mark as soon as 5 had been done would, in our opinion, have encouraged students to stop early. This tendency could have been overcome by awarding low grades for the exercises, but that ran contrary to the course philosophy of encouraging and rewarding mastery of the material. Alternatively, it could have been overcome by failing students outright for not doing all of their selected modules. That has the disadvantage of being inflexible and draconian, and was unlikely

to engender enthusiastic work on the last four modules. We were also keen to encourage students to attempt the P2 exercises, but did not want to allocate more than one ninth of the course grades on a (somewhat) subjective judgement about the quality of their P2 work.

A non-linear grading formula solved both problems neatly. Basically, we allocate one ninth of the marks to each of the practical modules attempted. However, the number of P2 practicals passed is used to weight the P1 grades. If no P2s are completed, then the P1s are weighted at two-thirds of their standard value. Three P2s must be completed to bring the P1 weighting back to 1. Assuming these weighted values (in the range 0-1) for P1 modules, our grading formula is:

$$Score \equiv \frac{100}{9} * \left[\left(\sum_{P1's} Grade_{P1} \right) * \frac{6}{9 - \#P2's} + \left(\sum_{P2's} Grade_{P2} \right) \right] \qquad \text{for 0-3 P2s}$$

$$Score \equiv \frac{100}{9} * \left[\left(\sum_{P1's} Grade_{P1} \right) + \left(\sum_{P2's} Grade_{P2} \right) \right] \qquad \text{otherwise}$$

The effect of the formula can be shown by graphing the maximum score possible against the number of practicals completed (Figure 5). The graphs are drawn assuming that students do their P1s first, and finish with P2s, which was usually the case. (If P2s were to be done first the grading formula would give conventional (linear) results.)

The system encourages students to complete the course by not giving them a pass mark until they have completed 6 of the 9 practicals. Once they reach that point, a high grade is clearly within grasp for a reasonably small amount of additional effort - ie: the marginal return on effort is high. Since it is the doing of P2s, rather than the grade achieved, which scales the P1 marks, we therefore avoid multiplying the effects of a small number of subjective grading decisions. For students who complete the course, each practical counts equally towards their final score.

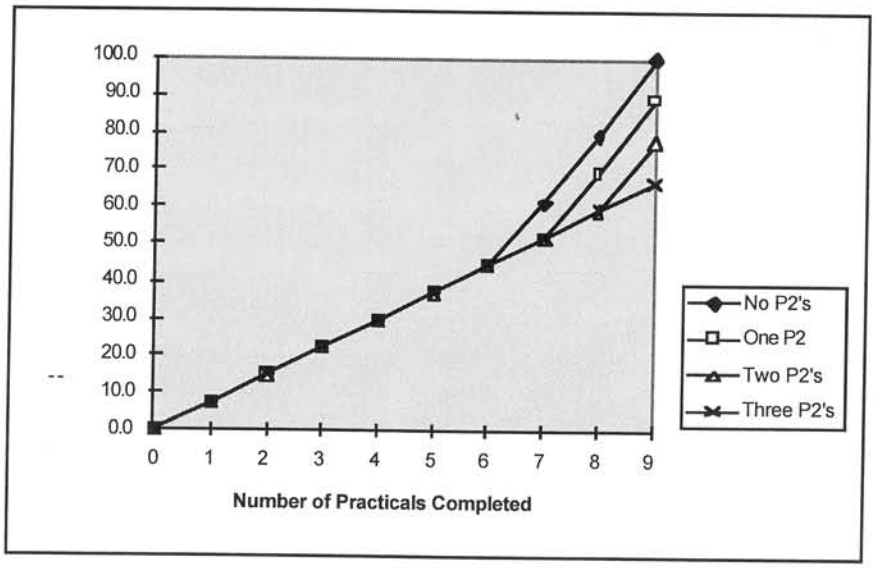

Figure 5. Graph of maximum possible score against number of practicals completed

Figure 6 presents the grade distribution for the 1995 B semester of The Computing Experience (the A semester also had a similar distribution). The marked shift in the distribution to the high end of the grading scale is typical of criterion-referenced assessment. Since students could work at their own pace on modules, they tended to continue plugging away until they met all of the objectives and received full marks for each module they attempted. Indeed, our suggestion that students having trouble with a particular piece of software should accept a low mark in the practical and move on to the next was almost uniformly rejected. The effect therefore was for students to earn a four for their P1s. Since the criteria for marking P2s was published and students could put in extra time in the computer labs to polish their work, high marks tended to be awarded for P2s as well.

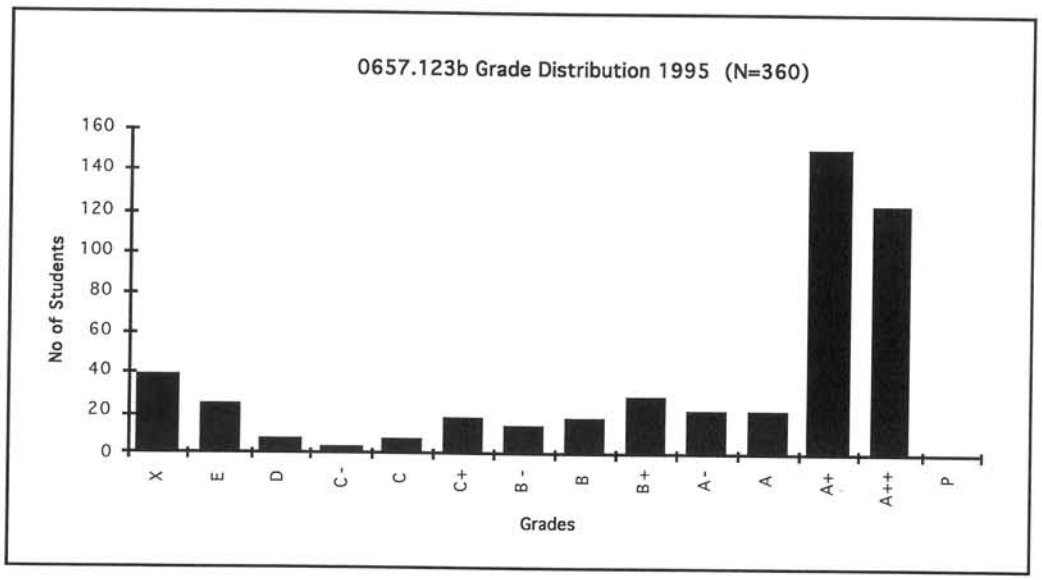

Figure 6. A final grade distribution for The Computing Experience

The nonlinear formula did in fact further spur students on in completing all of their selected practicals. Those who chose the 6 P1 / 3 P2 mixture, for example, found that after completing all 6 P1s their cumulative mark for the course was 44% — a failing mark. The very heavy weighting of the P2s encouraged the students to finish at least two of them, with the carrot of a high grade drawing most students to complete all their selected modules. Moreover, very few students deliberately chose P1/P2 mixtures that would not allow them a chance at an A. This last point took us by surprise, as we had assumed that there would be a significant minority of students who would opt to do as little as possible to achieve a bare pass to the course.

Finally, note that the percentage of students awarded an "X" has been nearly halved to 10.8%. The grades given in The Computing Experience reflect the trends that fewer students were informally dropping the course, those who stayed in the course set generally high goals for themselves, and most worked at the practicals until they achieved those goals. Moreover, the students could tailor the choice of practicals to their own interests and level of prior computing experience — an improvement over the uniform assignment topics required in the earlier Computer Applications course.

4. Reactions to the new grading scheme

Such a radical change in the class was accompanied by teething pains, of course: some new modules were too time consuming, essential software failed at inopportune moments, practical manuals contained errors or omissions, etc. Nonetheless, the student response to the new approach was generally good; The Computing Experience received high ratings in the course evaluation. Students appeared proud of the work they had done, and were pleased with their high grades. Most students seemed to feel that they deserved the marks that they achieved. Some of the faculty members in other schools of study were also enthusiastic about the course — to the extent that one Board of Studies praised it by a vote of acclamation, and commended the course for better meeting the needs and interests of the students.

The high marks attained by students was of grave concern in other quarters, however. One School threatened to refuse to allow its students to take the course, citing a concern that any class awarding such a large number of "A"s must be lacking in academic standards. Another objection voiced was that students taking the course would have an unfair advantage over peers choosing other classes, in that those in The Computing Experience would have a higher grade average. For pre-medical students, the difference in grades could be crucial in determining admittance to medical school.

5. Conclusions

In short, we felt that we had set up a grading scheme that encouraged hard work and eventual mastery of the material in most of the students. However, this criterion-referenced marking was engendering grades that were not in line with the generally norm-based assessment used elsewhere in the University.

Our scheme could describe in detail the learning objectives achieved by individual students, but could not provide a ranking of students as demanded by large segments of the University faculty. We found it ironic that when a normal curve of grades had been produced in earlier years, there had been little inquiry about how this curve had been produced and what precisely the students had learned. The mere fact that a normal distribution had been achieved had apparently been taken as evidence that the course contents were of an acceptable standard. Similarly, the skewed grade distribution for The Computing Experience was seen by some as proof that the course had no intellectual content.

We had anticipated that students would earn higher grades in The Computing Experience than in Computer Applications, but the extent of the grade increase took us by surprise. We had assumed that module completion would follow the distribution of practicals in Computer Applications: that a significant minority of students would opt to do less work, and to leave portions of the course uncompleted. Instead, the module completion distribution more closely resembled that of the graded assignments in Computer Applications: most students chose to do most of the work, and to continue working until a high grade could be assured. Hindsight suggests that we should have anticipated this outcome, since students work primarily to receive grades, and the modules in The Computing Experience were graded (like the assignments in Computer Applications).

This year, we are attempting to maintain the advantages achieved by our original formula (a lower rate of "X" grades and a high rate of module completion) while avoiding the main source of criticism for the course (the large number of high "A" grades awarded). Basically, we have reworked the grading scale so that the best grade that can achieved solely by completion of practicals is an "A". Higher grades — A+ and A++ — will be awarded to students whom we feel have performed exceptionally well on their P2s. While the nonlinear grading scheme was workable in the context of criterion-referenced assessment, the grade distribution it produced was not acceptable within the primarily norm-based University grading framework.

References

Bilimoria, D. (1995) "Modernism, postmodernism, and contemporary grading practices." *Journal of Management Education* 19(4), pp. 440-457.

Cross, L.H., Frary, R.B., and Weber, L.J. (1993) "College grading: achievement, attitudes, and effort." *College Teaching* 41(4), pp. 143-148.

Durm, M.W. (1993) "An A is not an A is not an A: a history of grading." *The Educational Forum* 57, pp. 294-297.

Echauz, J.R., and Vachtsevanos, G.J. (1995) "Fuzzy grading system." *IEEE Transactions on Education* 38(2), pp. 158-165.

Farley, B.L. (1995) "'A' is for average: the grading crisis in today's colleges." ERIC document ED384384.

Hinely, R.T. (1968) "An equal chance — a fantasy." In B.L. Turney (ed) *Catcher in the wrong: iconoclasts in education* (Itasca, IL, USA: Peacock Press), pp. 72-76.

Kuby,L. (1994). "My great grading experiment: motive and outcome." *Contemporary Education* 66(1), pp. 28-31.

Rowntree, D. (1987) *Assessing students: how shall we know them?* London: Kogan Page.

Wiggins, G. (1988) "Rational numbers: toward grading and scoring that help rather than harm learning." *American Educator*, Winter 1988, pp. 20-25 and 45-48.

Wood, L.A. (1994) "An unintended impact of one grading practice." *Urban Education* 29(2), pp. 188-201.

Wright, R.G. (1989) "Don't be a mean teacher." *Science Teacher* 56(1), pp. 38-41.

Wright, R.G. (1994) "Success for *all*: the median is the key." *Phi Delta Kappan*, 75(9), pp. 723-725.