The fallacy of "grading to the curve"

Keith McGuinness
School of Environmental & Life Sciences

*'"It is **not a symbol of rigor** to have grades fall into a 'normal' distribution; rather, it is **a symbol of failure** -- failure to teach well, failure to test well, and **failure to have any influence at all on the intellectual lives of students**."*[1] *[emphasis added]*

Statistical arguments

Because my statistical proclivities are well known, I will start with the statistical arguments. Suppose that I am about to teach SID569, *Analytical Critique of Icelandic Origami from a Postmodern Perspective,* for the first time to a class of three hundred students[2]. As this is a new subject, I decide to run a test in the first week to gauge students' initial knowledge of this important topic. I construct, and administer, a simple four option multiple choice quiz with twenty-five questions worth four marks each. Assuming that the students have no prior knowledge, and select answers completely at random, the average mark will be, obviously, about 25% and the spread of marks will resemble the distribution on the left in Figure 1 (the "Before Mark")[3]. This distribution is clearly not normal: it is skewed to the left and has a tail extending to the right. The distribution plotted is, in fact, binomial and *cannot be normal*, as it is truncated at zero (and also at 100, although that does not influence the "before" distribution). (A normal distribution, with a mean of 25, would have some students scoring less than zero; an impossibility here.)

After twelve weeks of intensive lessons, using the best technology and most advanced pedagogical methods, I administer and mark the end of semester exam: for simplicity, we will assume that it has the same format as the initial quiz. If we assume that my teaching has been effective, and students now have a 75% chance of making the correct choice on each question, the distribution of results will resemble that on the right of Figure 1 (the "After Mark"). This distribution is also clearly not normal: it is now skewed to the right, with a tail extending to the left (it is, in fact, the mirror image of the "Before Mark" distribution). Despite being clearly non-normal, the distribution actually resembles the spread of marks often seen[4].

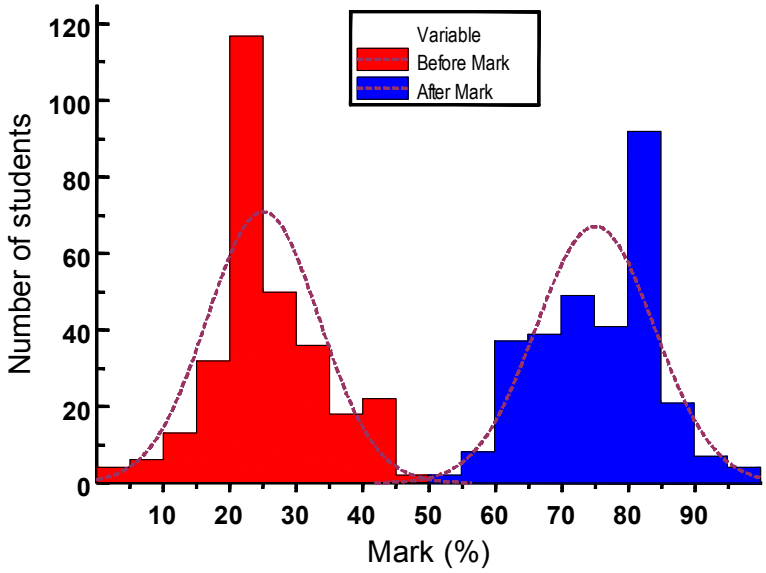

Figure 1. Distribution of marks in the hypothetical subject SID569 at the start of semester, left (red), and end of semester, right (blue). The curves overlaying the histograms are normal distributions, with appropriate means and standard deviations.

On the issue of grade distributions, Falkenberg (1996) states:

> *"The only way that the underlying distribution for students scores on the final exam in a college course can be normally distributed is if the students are a random sample from the population. If you only administer the test to students who have enrolled in the course or students who have completed the course,* ***the sample is clearly not random and a very different (and unknown) distribution may apply****. As soon as a faculty member* ***teaches the students something, their performance on the test is no longer random but rather a reflection of the outcome of the teaching-learning process."***[5]

On the history of "curving" grades, referring specifically to the discipline of psychology in the US he wrote:

> *"In the 1950's and 1960's it was popular for university faculty to 'curve' the grades in their courses. The assumption was that grades (like all the other psychological variables that were being studied by scientific psychology at the time) should be normally distributed. Early on, some faculty forced the grades into a normal distribution by applying a 'true curve'. Under this plan, regardless of the grades earned, the top 10% received A the next 20% were B's, the next 40% were C's, the next 20% were D's and the bottom 10% were F's,* ***regardless of the student's actual level of performance****. The* ***popularity of curving grades was relatively short lived****. Psychologists and psychometricians quickly realized that the* ***process was based on faulty assumptions and bad logic*** *and most had abandoned the process by the early to mid 1960's."*[op cit]

Abandoned in the early to mid 1960's? Perhaps not.

The effect of assigning grades

A further complication arises because, when results are reviewed, the emphasis is often on the spread of grades, rather than the distribution of marks. Although the underlying marks are a continuous variable, and could in some circumstances be approximately normally distributed, the grades are an ordinal (or ranked) variable and the application of the normal distribution in this situation is even more dubious. The most

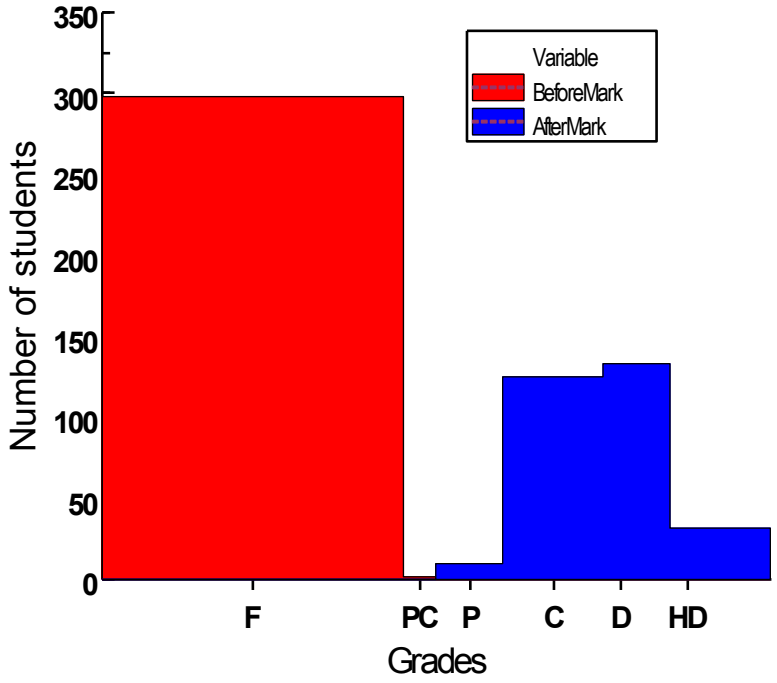

Figure 2. Distribution of results in SID569, with grades assigned according to the CDU assessment rules.

obvious reason is that the grade boundaries commonly applied in higher education are of varying widths: at CDU, an F is from 0 to 44, while a PC covers just the range 45 to 49. The effect of this is shown in Figure 2: arguably, assigning grades does not make the "after" distribution less normal, but it also certainly does not make it more normal.

Pedagogical arguments

Kohn (2002) states the basic argument clearly:

> *"A bell curve may sometimes – but only sometimes – describe the range of knowledge in a roomful of students at the beginning of a course. When it's over, though, any responsible educator hopes that the* ***results would skew drastically to the right****, meaning that* ***most students learned what they hadn't known before****."*[6]

Falkenberg (1996) notes that:

> *'While the statistical assumptions underlying curving were flawed,* ***an even greater flaw in the logic of statistical grade adjustment had been overlooked****.'[op cit]*

The 'flaw' he refers to lies in ignoring the wide range of factors likely to interact and influence student achievement. These factors include[7]:

1. the student's prior knowledge of the material;
2. the student's ability (unlikely to be normally distributed);
3. the student's motivation and effort;
4. the effectiveness of the instructional methods;
5. the quality of the instructional aids and resources, including assessment tasks; and
6. the ability of the faculty member engage students in the learning process.

There is no reason why these factors should always interact to give results which follow a particular distribution, or why the distributions in different units, taught by different staff, should be similar. (Distributions *similar* to the "after" plot in Figure 1 may occur in practice but this distribution is derived using assumptions which, while illustrative, are unrealistically simplistic.) Departures and differences are likely, especially in smaller classes (less than thirty to fifty) in which the performance of a relatively small number of students might greatly skew the results.

This does not mean that there should be no scrutiny or moderation of results or grades. First, 'grading to the curve' and moderating are two *very different processes*. The former is an attempt to force marks or grades to fit some largely arbitrary, and logically unjustifiable, distribution. Moderation, in contrast, is the process of ensuring that the marks and grades awarded accurately reflect student achievement on the assessment tasks. This should also involve checking that the task itself is appropriate, given the learning objectives of the unit. When units are team taught, moderation may also be required to maintain an even level of assessment across different staff and tasks. All of these are indispensable aspects of good teaching but none involves 'grading to the curve' or checking to see if "too many" distinctions and high distinctions have been awarded. Indeed, Falkenberg (1996), possibly tongue in cheek, recommends that we:

> *"****Identify faculty who grade on a curve and send them for remediation****. It is understandable that a first or second year faculty member would misjudge the level and ability of the students and write some tests that the students can't pass. But faculty who do not show substantial improvement in this area, will* ***probably need outside intervention and retraining to overcome the deficit in test preparation skills****."[op cit.]*

"Please explain"

> *"Grading on the curve makes learning a highly competitive activity in which students compete against one another* ***for the few scarce rewards (high grades) distributed by the teacher****."*[8]

> *"If a faculty member gives all A's give her/him a big raise. Any professor who can get all her/his students to master college algebra for example, should be rewarded for meritorious performance."*[5]

It is important to remember that much of the assessment in our science-based disciplines is criterion-based. The correct answer to the question "Is a fish an invertebrate?" is "No"[9]. If all the students in a unit, score 85% or better on an assessment task, then they should all be awarded high distinctions[10]. The suggestion that, if the percentage of HD's awarded exceeds some arbitrary value, there is necessarily a problem requiring investigation is unsound.

This is the current *paradox of quality improvement* in teaching in higher education. We are all encouraged to improve our teaching and learning practices. A necessary consequence of such improvement, assuming that there is no change in the nature of the students, is that the percentage of high grades awarded will increase. Indeed, if, in fact, there is *no such increase*, it is difficult to argue that *teaching practices have actually improved.* It is critical here to remember that the primary task of the teacher is to help the students achieve the stated learning objectives. The assessment tasks, if properly constructed and evaluated, measure how well the students have progressed towards mastering these learning objectives. If teaching and learning practices have improved in *meaningful ways*, then marks and grades *must rise*[11].

Such an increase in marks and grades may, however, be interpreted as an example of "grade inflation". The teacher will then often be required to justify the awarding of "too many" high grades and to demonstrate that they have not "gone soft" to gain the approval of the students. The suggestion may even be made that the assessment tasks should be made more difficult, simply to reduce the proportion of high grades. Unless, however, there is clear evidence of inappropriate practice, it would be better to congratulate the teacher, and ask them to describe their changed practices, so that other staff and students could benefit.

> *"Grade distributions could become more negatively skewed [towards higher grades] because the faculty have developed or learned more effective teaching methods and are more effectively motivating students to learn the material. In this case the students will learn more and hence grades will be higher.* ***This is the good kind of grade inflation. This is the kind that universities should encourage.*** *Higher education has a long and noble history.* ***Surely, we have developed more effective methods of teaching than were available half a century ago****."*[5] *[emphasis added]*

Notes and references

[1] Ohmer Milton, Howard R. Pollio, and James A. Eison (1986) *Making Sense of College Grades*. Jossey-Bass: San Francisco.

[2] I am using 300 because the largest SELS unit has about this many students and "grading to the curve" requirements are usually applied to large classes.

[3] Note that I have generated, and plotted, actual marks, rather than display the curve of the theoretical distribution, to illustrate the "lumpiness" that can occur even with very large classes.

[4] I am not suggesting that these distributions will generally be binomial in shape: it happens in this case because of the way in which I constructed the marks. Also, in practice, we often see a "clump" of low marks, and F grades, which are from students who stopped participating in the subject but did not formally withdraw.

[5] Steve Falkenberg (1996) Grade inflation. Online at: http://people.eku.edu/falkenbergs/grdinfla.htm#Statistical%20concepts. Accessed July 2, 2010.

[6] Alfie Kohn (2002) The Dangerous Myth of Grade Inflation. *The chronicle of higher education*, November 8, vol. 49, No. 11, p. B7. Online at: http://www.alfiekohn.org/teaching/gi.htm

[7] List is modified from Falkenberg (1996): op cit.

[8] Thomas R. Guskey (1996) *Communicating Student Learning: The 1996 ASCD Yearbook*. ASCD: Alexandria, p. 18.

[9] I am not suggesting that rote regurgitation of "facts" is necessarily good assessment practice, although there will always be information that students should remember, and be able to repeat and use, to be competent in their discipline.

[10] This does require that the assessment task is set and marked at an appropriate level. Falkenberg (1996) continues: "Of course this assumes than the grades reflect mastery of course content and that appropriate standards have been established and appropriate assessment methods have been implemented."

[11] Marks and grades are, of course, not the only measures of improved teaching. Increased retention and completion rates may also indicate improved practices but better learning should result in better grades.