



© Lucian Milasan | Dreamstime.com

Editorial

The question facing voters on 23 June is ‘Should the [UK] remain a member of the [EU] or leave the [EU]?’ with corresponding answers ‘Remain ...’ and ‘Leave ...’. This format is well designed: clear, concise and impartial. In contrast, regional referendums on Welsh devolution (2011) and Scottish independence (2014) had draft questions ‘Do you agree ...’, which would have led voters towards apparently favoured positions had they not been changed before implementation. However, their simple answers of ‘Yes’ and ‘No’ could still have introduced bias as affirmation is generally deemed more pleasing than negation. These were also the options for nationwide referendums on the European Community (1975) and alternative voting (2011). The current referendum avoids these problems and has fair campaign rules, though extensive feedback arising from media reports, opinion polls and social networks will surely influence the outcome.

Other questions and answers are prominent at this time of year, with examination setting, invigilation, marking and board meetings. Such assessments pose many puzzles of interest, particularly in question formats, marking criteria and grade adjustments. As the work submitted by pupils at schools and colleges is collated for marking nationally, huge cohorts are considered and authorities can set grade boundaries to ensure temporal continuity and achieve political goals. However, universities design their own syllabi and set their own examinations. As a result, assessments usually involve small cohorts and grading systems vary among departments and institutions.

When I began my academic career, we aimed to set examinations with long-term averages of 55% such that 10% of students achieved first-class marks of 70% or more and 10% achieved unclassified marks of less than 40%. Now we rely upon the skills

and experience of teachers, internal moderators and external examiners to maintain standards, while exam boards reserve the right to adjust module marks for consistency with other modules. This process works well, except for the discontinuity of deciding whether to scale a module’s marks and the inability of algorithmic scaling to distinguish between students’ abilities and assessment standards. Three years ago, I taught a final-year maths module to a group of three intelligent students who achieved a high average of 82%. Rigorous moderation confirmed these marks, whereas algorithmic scaling would assume that the questions were too easy or the marking too generous.

Adjustments preserve rank order and are usually applied only in exceptional cases when there is evidence that an assessment was too easy or too difficult. Various algorithms for modifying the original marks $x_i \in [0, 100]$ are then used by assessment boards to generate revised marks $y_i \in [0, 100]$ including the linear transformation

$$y_i = a + (b - a) \frac{x_i - a_o}{b_o - a_o}, \quad (1)$$

which fixes the desired minimum mark a and maximum mark b in terms of their observed equivalents, a_o and b_o . Equation (1) can be modified to fix the mean and standard deviation instead, or extended to fix the median and extremes using a piecewise linear transformation. I prefer a simpler approach for scaling original marks to generate revised marks, which is based on a factor $c \in [0, 1]$ and involves setting

$$y_i = cx_i \quad (2)$$

for an exam that was too easy and

$$y_i = 100 - c(100 - x_i) \quad (3)$$

for an exam that was too difficult. With $c = 0.8$, for example, this approach maps $[0, 100] \rightarrow [0, 80]$ for equation (2) and $[0, 100] \rightarrow [20, 100]$ for equation (3). Rather than specifying c to fix the maximum or minimum, we could achieve a desired mean (or quantile) m by setting $c = m/m_o$ in equation (2) and $c = (100 - m)/(100 - m_o)$ in equation (3), where m_o is the observed mean (or quantile).

Another method might be relevant if systematic scaling were required. Purely for internal monitoring purposes, the maths department at Keele University employs an effective algorithm, which is a modification of leave-one-out cross validation. Each student's mark for each module is predicted using his or her mean from other modules and this is then compared with the observed mark. The average discrepancy for each module is then calculated as a suitable measure for comparing the setting and marking standards of all modules, both core and optional. I have not seen this approach used elsewhere, though it would surely be appropriate if transformations were adopted routinely.

Different problems emerge when regulations include discrete and continuous criteria for degree classifications. Students are often required to pass all final-year modules and are then classified according to their overall means. To illustrate the problems that can arise through combining minima and averages, consider the module marks (%) of three fictional students:

- Ron: 40, 40, 40, 40, 40, 40
- Harry: 70, 70, 70, 70, 70, 70
- Hermione: 100, 100, 100, 100, 100, 39

Under typical university degree classification systems, Ron earns third-class honours with an average of 40% and Harry earns first-class honours with an average of 70%. However, Hermione receives an ordinary degree as she failed a module, despite an outstanding average of about 90% that utterly outshines all individual marks achieved by her classmates. How is this fair?

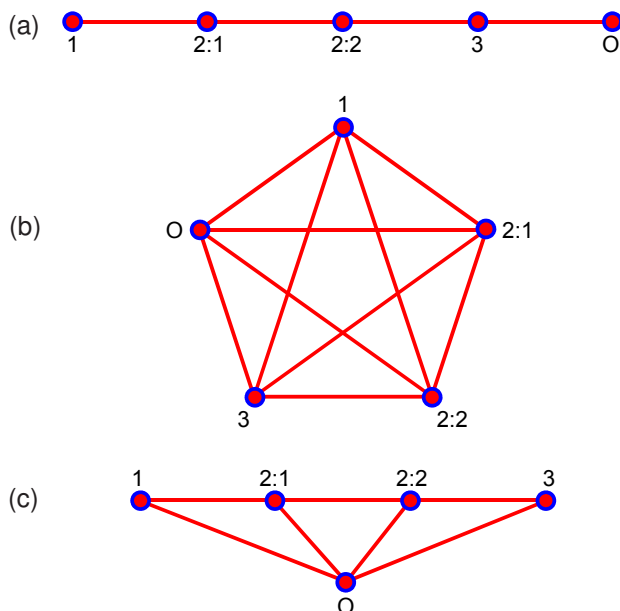


Figure 1: (a) Presumed, (b) typical and (c) compromised degree classification boundary links.

Most exam boards would treat Hermione's 39% as though it were 40%, so 1 mark out of 600 decides whether she achieves a first or is unclassified. But what if her 39% were 20% (average 87%) or 0% (average 83%) instead? Rules that combine thresholds and

averages can clearly generate bizarre results. Such oddities can occur at all grade boundaries, so although we might presume a sequential progression through degree classifications (Figure 1a), the reality can be rather peculiar (Figure 1b). If a reasonable standard is required in all modules, then a compromise would retain only the threshold between honours and ordinary degrees. Classification is then based on the student's average mark, conditional upon achieving at least 40% in each module (Figure 1c). Students must then be made fully aware of the importance of this threshold.

Now consider a multiple-choice exam of 20 questions, each with four possible answers. If a student answers all questions randomly, the probability distribution for the number X of correct answers is binomial $\text{Bi}(20, 1/4)$ and he or she can expect to answer $E(X) = 20 \times (1/4) = 5$ questions correctly by pure chance. Moreover, the student will achieve a pass mark of at least 40% with probability

$$P(X \geq 8) = \sum_{x=8}^{20} \binom{20}{x} \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{20-x} \approx 0.10,$$

corresponding to a 10% chance of passing despite total ignorance of the subject matter. Suppose that a student scores $X = x$ correct answers and we wish to ascertain his or her true ability, as measured by the unknown number of questions Y for which the student knew the correct answers. Specifically, we must determine the conditional distribution of Y given $X = x$.

We assume a prior distribution for Y of the form $\text{Bi}(20, 11/20)$ so that the mean corresponds with 55% as suggested earlier. From above, the conditional distribution for $X - Y$ given $Y = y$ is binomial $\text{Bi}(20 - y, 1/4)$. Applying Bayes' theorem then gives

$$p(y|x) \propto p(x|y)p(y) = \left\{ \binom{20-y}{x-y} \left(\frac{1}{4}\right)^{x-y} \left(\frac{3}{4}\right)^{20-x} \right\} \left\{ \binom{20}{y} \left(\frac{11}{20}\right)^y \left(\frac{9}{20}\right)^{20-y} \right\}$$

for $y \in \{0, \dots, x\}$ with $x \in \{0, \dots, 20\}$. A little bit of algebraic rearrangement reveals that the posterior distribution for Y given $X = x$ is also binomial, of the form $\text{Bi}(x, 44/53)$. We can now use $p(y|x)$ to evaluate quantities of interest, such as the probability that a student with 15 correct answers out of 20 questions is of first-class standard. This is given by

$$P(Y \geq 14|X = 15) = p(14|15) + p(15|15) \approx 0.25,$$

so, although this student has a clear first-class mark of 75%, he or she is three times more likely to be of second-class or lower standard. This generosity of multiple-choice tests contributes substantially to grade inflation.

Meanwhile, the Editorial Board has collated a wide range of fascinating articles by notable authors for this issue, including a film review of *The Man Who New Infinity* and an article on pedagogy. Other topics considered are book identifiers, science fairs, historical notes, numerical rounding, Kalman filters and tennis scheduling, so there is plenty of variety that should appeal to all. Finally, please complete our readers' survey if you have not already done so, to help us maintain and improve the standards and content of *Mathematics Today* – <http://tinyurl.com/MT16-survey>.

David F. Percy CMath CSci FIMA