Norm Referencing or Criterion Referencing?

Judith Pollard, The University of Adelaide
judith.pollard@adelaide.edu.au

Arguments for norm referencing

I'm probably talking to a room full of people who are already convinced that assessment should be criterion referenced rather than norm referenced. In fact most of my colleagues use criterion referencing in the assessment of second and third year subjects. But, they are likely to revert to norm referencing when dealing with first year classes, particularly when those classes are large and when the exams comprise many questions, or parts of questions, worth a few marks each.

Similar distributions

One argument for norm referencing goes something like this:

In large classes, the distribution of students is likely to be similar from one year to the next, so the distribution of grades should also be similar. On this basis, the same percentage of students is awarded a given grade from year to year. There are many problems with this approach:

- If there is an improvement in teaching and learning it cannot be reflected in improved outcomes.
- Students will be more reluctant to work co-operatively if they see that someone else can get ahead only by pushing others back.
- Variations do occur in the ability and interest of different cohorts of students, and therefore in their performance.

Gaps in the distribution

A related but slightly different approach is taken by those who say that the marking cannot be done with a precision better than 2 or 3 marks, so it is not reasonable to differentiate between students whose marks differ by only 1. Therefore grade boundaries must be placed where there are gaps in the histogram of results.

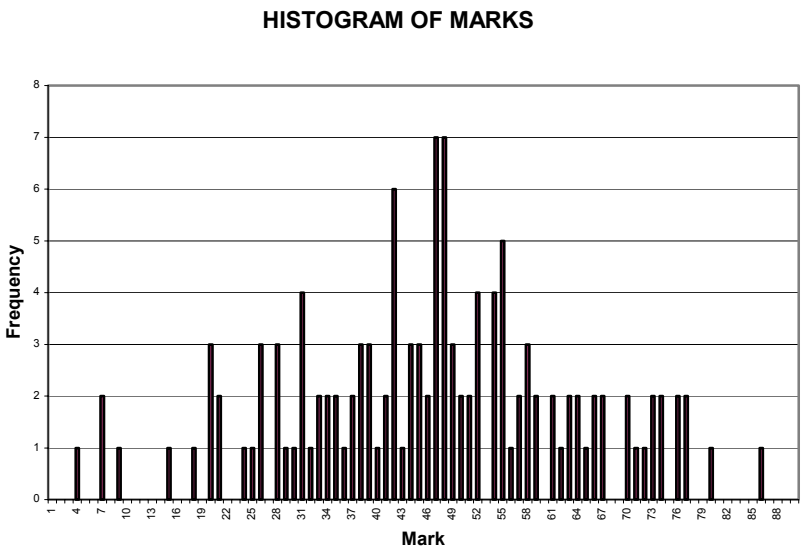

Figure 1: Histogram of results in an exam with a maximum mark of 90.

Figure 1 shows a histogram of results in an exam with a maximum mark of 90. Certainly, there are some gaps in the distribution, but what do they mean? What justification is there for linking any of them to grade boundaries?

Establishing criteria

Our department has adopted a set of criteria for grades in coursework. They are not perfect, but they work fairly well.

Description of grades

HIGH DISTINCTION (HD): Superior performance, showing security of knowledge, comprehensive understanding of the subject, initiative and originality.	≥ 85%
DISTINCTION (D): Superior performance, showing comprehensive understanding of the subject, with some evidence of initiative and originality.	≥ 75%
CREDIT (C): Demonstrating sound knowledge of concepts and principles, and the ability to apply them in standard situations over a broad range of topics.	≥ 65%
PASS, Division 1 (P1): Basic understanding, with knowledge of principles and concepts at least adequate to communicate intelligently in the discipline and to serve as a basis for further study, but with definite deficiencies.	≥ 55%
PASS, Division 2 (P2): Basic understanding, with knowledge of principles and concepts at least adequate to communicate intelligently in the discipline, but with substantial deficiencies which may make success in subsequent courses unlikely.	≥ 50%
CONCEDED PASS (CP): Unsatisfactory performance but with enough knowledge of concepts and principles that combining this result with performance at Pass level in another component would have given an overall grade of Pass.	≥ 45%
FAIL (F): Unsatisfactory performance with fragmentary knowledge of concepts and principles, or failure to complete the subject requirements.	

Different subjects have different grading schemes, so not all of these criteria are used in a given subject.

Matching criteria to performance

Preliminary grade boundaries

When the lecturer has marked enough scripts to formulate a reliable marking scheme, s/he makes two decisions:

Which marks in each question should a student earn, to just satisfy the criterion for a Pass Division 1? Call the sum of these marks A.

Which marks in each question can a student earn by demonstrating comprehensive understanding, initiative and originality (the criteria for a Distinction)? Call the sum of these marks B.

If the maximum mark for the exam is T, these values are used to get a ballpark figure for two grade boundaries:

$$P1 = 0.9\ A.$$

The factor of 0.9 is included to allow for the fact that in the stress of exams, students are bound to make silly mistakes, so we would like to leave a bit of slack.

$$D = T - B.$$

This means that a student cannot be awarded a Distinction without earning **some** of the marks which demonstrate comprehensive understanding, initiative and originality.

Final grade boundaries

I've said these results give ballpark figures. When we come to decide grade boundaries, we look at the marked scripts, starting with those within 2 or 3 marks of the expected Pass Division 1 boundary. At this point, we are looking for evidence that the student has a basic knowledge of principles and concepts, and shows some ability to communicate using the language of physics. There is usually a clear demarcation between those students and students who have picked up marks without displaying such evidence.

In a similar way, we look at scripts around the Distinction boundary, to find students who have earned the marks we identified as demonstrating comprehensive understanding, initiative and originality.

Other grade boundaries are then fairly easy to set:

High Distinction is about 10 % above the Distinction boundary, where students demonstrate comprehensive understanding, initiative and originality in most sections of the exam.

Credit is usually midway between P1 and D, where students demonstrate basic understanding in all or most sections of the exam.

Finally, where applicable, the P2 boundary is set where we would permit students to count the subject towards their degree, but not to pursue physics studies any further.

As a matter of interest, the grade boundaries for the histogram in figure 1 were set as follows: P2 – 39; P1 – 44; C – 55.5; D – 69; HD - 76

Combining components into a final result

Commonly, a First Year subject comprises two semesters of coursework (assessed mainly by examination, with some continuous assessment component) and practical work. If each component of the course has been correctly graded, the final result should be the weighted sum of individual components. We apply this procedure, then look at the outcome and sometimes make adjustments. For example:

We are reluctant to award P1 to a student with P2 in both coursework components, whose final grade boosted by a very good practical mark. We look again at the exam scripts to decide whether the student displays enough grasp of physics to have a reasonable chance of success in second year.

At the upper level, if a student has Distinction in 2 components, but an overall mark just below the D boundary, the exam scripts would be reviewed to see if a Distinction can be justified.

Using the criteria for different year levels

The criteria are interpreted within the context of the subject matter and the specific objectives of each subject. The differences in satisfying the criteria at different year levels have not been made explicit, but do not seem to present significant problems to lecturers.

One concern does arise, though, in relation to First Year students. It takes time for some students to develop study habits appropriate to tertiary study, yet at present we interpret the criteria similarly for First Semester and Second Semester exams. Many students do poorly in First Semester, but respond to the message, change their approach to study and pass the year.

We will need to review the way the criteria are applied if we semesterize all our First Year subjects.

Trouble-shooting

Too much scaling needed

Care needs to be taken in setting the examination, to ensure that an appropriate number of marks can be earned by students with basic understanding, and that there is the right amount of opportunity to demonstrate security of knowledge, initiative and originality.

There are three stages at which these parameters can be managed.

1. In setting the exam, we try to have 60% of the marks available for basic understanding, and 15 to 20% available for initiative and originality. This allows students to do something silly about 10% of the time, and still have grade boundaries around 50% for a Pass, and 75% for a Distinction.
2. Sometimes, the lecturer misjudges the difficulty of some of the questions. To compensate for these inadequacies, the s/he can be generous or tough in the marking scheme. After marking the first 20 or so scripts, the lecturer may realize that the marking scheme is too generous or (more likely) not generous enough. Adjustments can be made at that stage, so that the grade boundaries will correspond to appropriate exam marks.
3. This procedure comes unstuck if the first papers marked are not representative of the class as a whole. Then the grade boundaries might be well away from the "standard" boundaries. If it is too expensive or time-consuming to remark the papers using a new marking scheme, we have to live with a large amount of scaling, though this tends to reduce credibility.

We provide an explanation to our students of the criteria for determining the grade boundaries, and remind them about it when we show how exam marks have been converted to grades.

Mark order does not match criteria

It happens sometimes that a student who satisfies the criteria for D has a lower mark than a student who doesn't. Usually, a check of the marking reveals that one of the students has been marked too harshly, or the other too generously, and the discrepancy in mark order can be corrected.

Conclusion

It is possible to develop a set of verbal descriptors for assessment grades which can be applied across subjects and across year levels in a discipline. If it is adopted by the whole department, it can have a significant effect in changing the culture from norm referenced to criterion referenced assessment.

Discussion questions

1. In your department, how common is norm referencing?
2. What do you see as the major obstacles to changing the assessment culture of a Department/School/Faculty?
3. What changes would you make to the Descriptions of Grades in applying them in your discipline?
4. In a situation where student learning improves, for example if the standard of the cohort and/or the teaching methods improve, should the assessment standards change?