Grading Systems

The subject of grading is rarely discussed among faculty members, except perhaps for the occasional debate about grade inflation. But many teachers privately confess that grading is one of the most difficult and least understood elements of their job. Often, professors have little confidence that their grading systems accurately discriminate between different levels of achievement and they differ widely on the components that should constitute a final grade. As a result, grading standards and criteria are so idiosyncratic that an "A" from one teacher may be the equivalent of a "C" from another. Part of the problem with grading arises from the fallibility of the tests and assignments used to measure student performance. The three previous FYC's focused on ways to improve assessment techniques; in this article, we will survey several different methods for calculating final grades and point out their strengths and weaknesses.

Grading and Feedback

First, it helps to make a distinction between grading and other forms of feedback. A grade is a "certification of competence" that should reflect, as accurately as possible, a student's performance in a course. If this goal is achieved, then grades will have the same value from semester to semester and from year to year. Trouble arises when we include grading components that are difficult to measure accurately (such as effort or participation) because these elements reduce the strength of the relationship between grades and academic achievement. Furthermore, when we use grades for reward or punishment, give extra credit for additional work, or grade on attendance, we contaminate the meaning of grades and reinforce the students' belief that a course grade has less to do with academic performance than with fulfillment of arbitrary requirements.

Of course, we must give students feedback in many of these areas of behavior, but using the grading system to convey this assessment is inappropriate. Moreover, we often complain that students are excessively grade-oriented, but by attaching a grade value to every aspect of student performance we actually reinforce our students' preoccupation with grades. Teachers should avoid using grades as incentives for performance and seek out non-graded methods for motivating students. For example, verbal "rewards" in class, individual conferences, and written critiques can provide positive and negative feedback without contaminating the grading system.

Elements of a Grading System

A good grading system must meet three criteria: (1) it should accurately reflect differences in student performance, (2) it should be clear to students so they can chart their own progress, and (3) it should be fair. Performance can be defined either in relative or absolute terms (comparing students with each other or measuring their achievement against a set scale), and each system has its defenders. But whichever grading scheme you use, students should be able to calculate (at least roughly) how they are doing in the course at any point in the semester. Some relative grading schemes make it impossible for students to estimate their final grades because the cutoff points in the final distribution are not determined until the end of the course. A complete description of the grading system should appear in the course syllabus, including the amount of credit for each assignment, how the final grades will be calculated, and the grade equivalents for the final scores. Also, students should perceive the grading system as fair and equitable, rewarding them proportionately for their achievements. From the standpoint of measurement, many different kinds of assignments, spread over the entire semester provide a fairer estimate of student learning than one or two large tests or papers.

Relative and Absolute Grading Systems

Relative (norm-referenced) grading systems are probably the most widespread in higher education. In

and poor students in "low achievement" classes may unfairly benefit. Also, the method does require some knowledge of statistics and the mathematical tranformations involved so that you are not "working blind."

Note: Some teachers use a related method to transform each set of raw test scores to standardized scores before averaging, but prefer to use another method for determining the final grades. Averaging tests with different means and different standard deviations is somewhat akin to adding apples and oranges, and raw scores cannot be weighted or averaged without introducing a bias. Transformation to standard scores adjusts for differences in means and standard deviations and thereby preserves the mathematical integrity of each score. "T scores," which have a mean of 50 and a standard deviation of 10, are often used for this purpose, and spreadsheets can be programmed to make the transformations automatically. Of course, you must explain to students how their scores are being transformed so they won't be confused about their averages.

Another relative grading scheme is the "gap method," but it is difficult to defend on the basis of statistics or measurement theory. In this method, students' total course scores are arranged in ascending order and the teacher looks for naturally-occurring gaps in the distribution of the scores. Unfortunately, the gaps may not reflect real achievement differences but simply chance occurrence, and they may not appear at reasonable points in the distribution. The primary advantage of the gap system is that there are fewer complaints about borderline grades, since students are unsophisticated about grading systems and will likely accept the gaps as proof of significant differences in performance.

Absolute Grading

Absolute (criterion-referenced) grading is based on the idea that grades should reflect mastery of specific knowledge and skills. The teacher sets the criteria for each grade, and all students who perform at a given level receive the same grade.

The simplest absolute grading scheme is "percent of total points possible." The teacher decides on the total number of points that a student could earn in the course and sets arbitrary achievement levels based on the total. The cutoff for "A" grades might be 90%, for "B's," 80%, and so forth, and it is assumed that a student who makes 83% knows 83% of the material. If every student scores above 90%, they will all receive "A's." Although this method does provide clear performance targets for students, there are several problems associated with it. First, the rationale for the cutoff scores is usually murky and often based on intuition rather than analysis. Second, the system is based on the assumption that the teacher can construct valid, reliable exams and assignments at consistent levels of difficulty throughout the course. Third, some teachers apply the same performance scale to every evaluation component,

a practice which does not take into account the variability of the assignments or adjust for particularly difficult or particularly easy assignments. Finally, some students may achieve a high number of points simply by doing well on many small, less important assignments.

Objective-based grading is perhaps the most sophisticated kind of absolute grading because the method attempts to equate grades with different *kinds* of performance. In all the grading systems reviewed above, the teacher assumes that students who receive good final grades have learned the more important material and mastered the more complex levels of thinking, but this assumption may not be valid. For example, students who do very well on objective exams and poorly on written assignments may earn a respectable final grade, but may not have mastered important intellectual skills that the teacher had in mind. The objective-based grading method takes into account both the amount of material students learn and the level of cognitive complexity they achieve.

To use objective-based grading, the teacher must first review the kinds of knowledge and skills that are *implicit* in the course and make them *explicit* as course objectives. You must identify two kinds of outcomes: minimal objectives and developmental objectives. Minimal objectives are statements of essential course outcomes and basic skills; developmental objectives reflect higher-order cognitive processes such as critical thinking, decision-making, and complex problem solving. Examples:

Minimum Essential Objectives
The student will be able to:
- describe different kinds of plasmids
- describe transposons
- explain how transposon mutagenesis works

Developmental Objectives
The student will be able to:
- work problems in bacterial genetics involving transformation, transduction, and conjugation
- design a protocol to clone a gene or obtain a particular mutant using transposons

It may be easier, at least initially, to measure achievement of minimal and developmental objectives using completely separate exams and assignments for each type of objective. This technique will simplify record-keeping and help you focus more sharply on the different kinds of tasks that are appropriate for assessing the two types of objectives. Test questions and exercises for minimal objectives are relatively easy to create because they assess basic knowledge and well-rehearsed skills. Measuring developmental outcomes is more difficult, for you must not only master the classification systems for complex thinking and reasoning skills but also must be able to devise assigments that measure these skills. Some writers suggest that "novelty" is one element common to higher-order learning tasks and

relative grading, students are in competition with one another for a limited number of grades in each category, and a student's grade is based on his or her relative position in the class. By contrast, absolute (criterion-referenced) systems use an unchanging standard of performance against which student performance is measured, so a student's grade is related to his or her achievement of particular levels of knowledge or skill. No grading system is foolproof, for the integrity of any system depends on the teacher's ability to devise valid and reliable measurements of student performance. Measurement error is therefore the greatest hindrance to effective grading.

Relative Grading

Relative grading is based on two assumptions: (1) one of the purposes of grading is to identify students who perform best against their peers and to weed out the unworthy, and (2) student performance more or less follows a normal distribution — the famous bell-shaped curve. Teachers who use relative grading point out that these systems correct for unanticipated problems (e.g. widespread absences due to a flu epidemic, tests that are too hard or too easy, or poor quality teaching) because the scale automatically moves up or down. Students like relative grading for much the same reason.
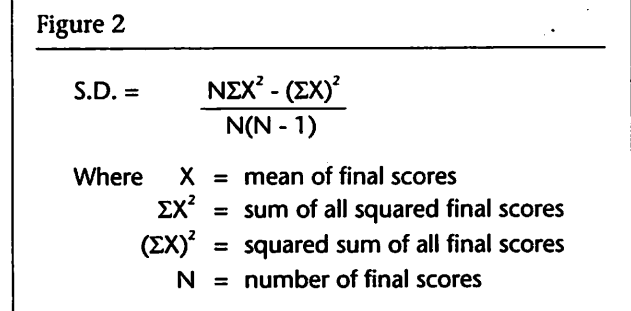
One of the most common relative grading systems is "grading on the curve." The use of the normal curve as a grading model is based on the discovery, earlier in this century, that IQ test scores over large populations approximate a normal distribution (Figure 1). Although it is true that the larger the class, the more likely that student performance will begin to look something like a normal curve, the assumption that performance is normally distributed is usually unjustified, even in large sections. In the first place, college students are a highly selected group, not representative of the general population with respect to background or intelligence. Second, we cannot be sure that our tests accurately measure student achievement — even standardized exams are suspect in this regard.

Figure 1
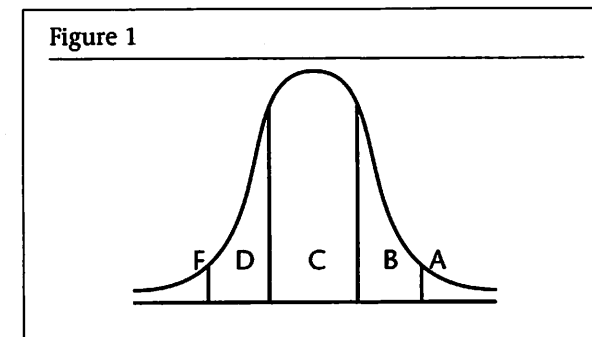
Fortunately, few teachers adhere to a strict normal distribution, since it will fail a fixed percentage of the class and award "A's" to a fixed percentage, without reference to the overall level of performance. Forcing students into this scale tends to wreak havoc with their motivation. Consequently, many people use a "skewed curve" in which the distribution is shifted upward slightly, resulting in fewer grades below "C" and more in the "B" category. However, few teachers base their modified curves on statistical principles or cumulative performance data; they simply select a distribution that "looks right." Typically, the rationale for grade cutoff points is based on tradition rather than on analysis of student performance over time. The major problem with any curve is that one cannot be sure that differences in performance are real or simply artifacts of the distribution — was the performance of the top 5 students who got "A's" substantially different from that of the 15 who received "B's?"

Statistically speaking, the soundest relative grading system is the standard deviation method. In this system student grades are based on their *distance* from the mean score for the class rather than on an arbitrary scale. To calculate the standard deviation, the teacher creates a frequency distribution of the final scores and identifies the mean (average) score. Using the formula in Figure 2, the standard deviation is computed so that cutoff points for each grade level can be determined (note: spreadsheets can be programmed to perform the math automatically).

Figure 2

$$S.D. = \frac{N\Sigma X^2 - (\Sigma X)^2}{N(N - 1)}$$

Where X = mean of final scores
ΣX^2 = sum of all squared final scores
$(\Sigma X)^2$ = squared sum of all final scores
N = number of final scores

Cutoff points for "C" grades range from one-half the standard deviation below the mean to one-half above. Adding one standard deviation to the upper "C" cutoff will yield the "A-B" cutoff point, and subtracting one standard deviation from the lower "C" cutoff will provide the "D-F" cutoff point (examples in Figure 3).

Figure 3

	Class I	Class II
Mean of final scores	79.2	60.76
Standard deviation	12.79	9.85
Upper "C" cutoff	85.6	65.69
Lower "C" cutoff	72.8	55.83
"A/B" cutoff	98.4	75.54
"D/F" cutoff	60.0	45.98

Although the standard-score method of computing grades is statistically superior to other relative grading methods, there are several cautions to keep in mind. As with other relative grading schemes, capable students in "high achievement" classes may be unfairly penalized

therefore assignments that require students to apply their thinking skills in new ways or in new situations will test complex reasoning.

If you develop tests and exercises that accurately assess both kinds of objectives, you can set performance standards and grade equivalents on a scale like this one:

Grade	Performance Standard	
	Essential Objectives	*Developmental Objectives*
A	90% or more	85% or more
B	90% or more	75 to 84%
C	80% or more	60 to 74%
D	80% or more	50 to 59%
F	less than 80%	less than 50%

In this example, in order for students to pass the course they must master at least 80% of the minimum essential objectives and 50% of the developmental objectives. Obviously, setting these cutoff points must be done carefully, taking into account the difficulty of the tests and assignments and student performance in previous classes or other sections of the same course. If using this kind of scale seems too difficult, you could use the "total points possible" system instead. By awarding more points for tests and assignments on higher-level objectives and fewer points for tasks on less important objectives, you would still reap some of the advantages of the objective-based method.

No single grading system will be appropriate for all courses at all times, and teachers must be sensitive to differences in students and subject matter when choosing a grading system. It takes time to develop realistic expectations about student performance, and the best

teachers constantly re-examine their grading assumptions to verify that their systems are valid. Finally, the accuracy of any grading system is dependent upon the validity and reliability of the measures we use to assess student performance, so improving the quality of exams and course assignments will improve the accuracy of the final grades.

Bibliography

Cheshier, S. R. (1975) Assigning grades more fairly. *Engineering Education, 65, 2.*

Ebel, R. & Frisbie, D. A. (1986) *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Gronlund, N. E. (1973) *Preparing criterion-referenced tests for classroom instruction.* New York: Macmillan.

Saunders, P. (1978) Instructional objectives. In P. Saunders, A. L. Welsh, and W. L. Hansen (Eds.) *Resource manual for teacher training programs in Economics.* New York: Joint Council on Economic Education.

Svinicki, M. (n.d.) Evaluating and grading students. In *Handbook for new faculty.* Austin: Center for Teaching Effectiveness, University of Texas at Austin.

Terwillinger, J. S. (1989) Classroom standard setting and grading practices. *Educational Measurement: Issues and Practice, 8, 2.*

Wright, D. L. (1987) Grading: Measuring students' achievement in a course. *Teaching at UNL.* Lincoln, NE: Teaching and Learning Center, The University of Nebraska-Lincoln.

CTL **Center for Teaching and Learning** 919-966-1289
CB #3470, 316 Wilson Library
Chapel Hill, N.C. 27599-3470