IDEA Paper No. 19

center for FACULTY EVALUATION & DEVELOPMENT
Division of Continuing Education
Kansas State University

January, 1988

Improving College Grading

by
Gerald S. Hanna and William E. Cashin
Kansas State University

> A grade is "an inadequate report of an inaccurate judgment by a biased and variable judge of the extent to which a student has attained an undefined level of mastery of an unknown proportion of an indefinite material." (Dressel, 1983, p. 12)

It is well known that the level of achievement sufficient for an A in one section of a college course taught by one instructor may well yield a C in a section taught by another instructor. It clearly is undesirable for a student's grade to be an artifact of who taught the course and/or what other students were enrolled in the section. Ideally, grades should be pure measures of achievement, uncontaminated by extraneous factors.

This IDEA Paper will first focus on the advantages and disadvantages of two common grading methods. From an analysis of the virtues and vices of these prototypic methods, a set of goals or criteria for grading systems will be derived. Next, the concept of anchor measures will be presented. Finally, three illustrative grading systems will be presented and evaluated by use of the above criteria.

Portions of the following discussion are based on the rationale presented in IDEA Paper No. 18 (Hanna & Cashin, 1987). Readers who have not studied this earlier IDEA Paper are encouraged to do so before proceeding.

Two Prototypic Grading Systems

There are two prevalent methods by which American college and university instructors assign student grades—"percentage grading" and "grading on a curve." There are, to be sure, other methods, but they are practiced by relatively small minorities of college teachers. McKeachie (1986) and Terwilliger (1971) discussed several of these less common grading methods and the problems associated with them.

We shall focus on the two most widely practiced systems. Although there are numerous variants of these two major methods, most instructors fall relatively neatly into the camp that practices some variety of percentage grading or the camp that, by one means or another, grades on class curves.

Percentage Grading Systems

Percentage grading systems are perhaps the most popular methods of assigning college grades. The instructor usually announces some "absolute standards" early in a course in the form of the percent of the possible points that must be earned in order to receive each grade. Advocates believe that such systems have the virtue of giving students advance notice of what they have to do in order to earn various grades.

This indeed would be desirable. But does an announcement that "you will need to average at least 90% to earn an A in this course" really communicate what students have to learn or how hard they will have to study? Does it define the domain of course content? Does it specify the difficulty of the tests on which they have to maintain this percentage? It does not!

It will be recalled from IDEA Paper No. 18 that most college level course work falls clearly into Track C. Such large, open, vaguely described content domains do not enable meaningful interpretation of student performance in terms of either raw scores or percent scores. To be useful, a domain-referenced (i.e., criterion-referenced) statement must relate to a content domain that is very **clearly specified**. "The domain definition should be sufficiently detailed and delimiting to show clearly what facets of behavior are included and what facets are excluded in the domain" (American Psychological Association, 1985, p. 26). "The fruitfulness of the orientation in terms of domain mastery depends on the possibility of defining a domain clearly and incisively, so that the range of performance that lies within the domain can be fully specified and agreed on" (Thorndike, 1982, p. 2).

An important consequence of the size, openness, and lack of incisive description that characterizes most college level test domains is **uncontrolled test item difficulty**. An instructor can (intentionally or inadvertently) develop a test on which no student is likely to obtain even 75% of the possible points. Similarly, an instructor can create a test on which most students can attain at least 90%. Raw scores and percent scores are not only functions of student achievement, but are also **artifacts of test difficulty**. For this reason, advance information concerning "absolute standards" such as, "you must average at least 86% in this class to earn a B" creates an **illusion** of informative clarity; **it really tells nothing**.

Such statements tell nothing, that is, unless there is an implicit understanding concerning test and assignment difficulty. Difficulty, of course, is inherently norm referenced, not domain referenced. Thus, for a percentage grading system to convey information, it must violate its intrinsic domain-referenced nature and be rendered covertly norm referenced.

As testimony to the truth of this strong claim, consider what happens when a test turns out to be too difficult. Suppose an instructor uses percentage grading in a class that has previously

evidenced reasonable competence. Now, on a particular exam, the top person receives only an 80%. Does the instructor give all low grades? Often not. Instead most teachers who find themselves in this predicament do one or another form of violence to the meaning of percentage grading. A few void the test. More go ahead and count it, but "rig" the next one to be very easy in order to compensate for the hard one. Many engage in such "eccentric" arithmetic practices as counting each of a hard test's 40 items 3% rather than 2.5%, while others might give everybody a 15% bonus. Some norm reference the percentage grades to the top person in the class (80% in this example); e.g., a student who originally received a score of 60% would be elevated to 75% (60 divided by 80). While norm referencing makes sense, one could not do worse than to select a single individual for the reference "group"!

None of these "adjustments" are compatible with the basic rationale of domain-referenced, percentage grading. Rather, they are ways instructors find out of the messes that a fundamentally illogical system gets them into. It would be far better to avoid these predicaments in the first place.

In content domains suitable for most college courses, the subject matter is variable in difficulty; e.g., one could test knowledge, understanding, or application of given facts or principles with relatively easy questions or with relatively difficult ones. Large, open, ill-defined content domains do not tie down test question difficulty. Thus, interpreting student performance in terms of either raw or percent scores—domain referencing—cannot be meaningfully achieved. (Space constraints preclude full development of this idea. See Hanna and Cashin, 1987, for somewhat more detailed discussion or Hanna, 1985, for a much longer treatment.)

Since domain referencing is not feasible with the kind of content appropriate for most college courses, it is appropriate to seek a grading system that is norm referenced.

Class-Curve Grading Systems

Instructors who use class-curve grading often recognize that domain referencing is not meaningful in the context of college grading and that norm referencing is the only viable alternative. (They may also recognize some or all of the problems of class-curve grading discussed in this section, but these may seem to be the lesser evil.)

Like those who use percentage grading, instructors who grade on class curves usually value advance notice to students regarding what is required to receive various grades. They seek this by such statements as, "To receive an A, you must be among the top 20% of the people in this section in total number of points at the end of the semester." But does such an announcement really inform students of how hard they will have to work in order to receive a given grade? No! Not unless they have prior knowledge of the academic capabilities of others who happen to be enrolled in the particular section.

One section of a course often contains better students than another. Persons in sections having many low achieving students can more easily rise to the top than can students in better sections. With sections enrolling fewer than 15 or 20, this section-to-section sampling error will ordinarily **make a one- grade difference for several students**; e.g., a student who is in the A stratum in the poor section might make only the B stratum in a more typical section. Even in groups of 30 to 50, sampling error can make a one-grade difference to a few students. The concern is abated only in very large sections enrolling several hundred students.

We have an apparent paradox. On the one hand, norm referencing is the only logical foundation upon which to base grades in typical college courses (cf. Hanna & Cashin, 1987). On the other hand, the only method of achieving norm referencing of which most instructors are aware—grading on a class curve—is ordinarily unsatisfactory because it introduces instability arising from small samples. Fortunately, there is a way out. But let us first identify the other major problems associated with class-curve grading.

Class-curve grading forces students to compete with one another **for grades**. But learning is not inherently competitive. (We have no universal objection to competition. Where there can be only one first chair clarinetist in the orchestra, competition seems inherent. If only one supervisor is needed for an assembly line, again competition makes sense. But there is no set fraction of students who can do A-level work in a class.) There is no reason why one student's success in learning must predispose others to less success. Yet in class-curve grading, there is a "preordination" of a section's grade distribution, **regardless of student learning**. This causes students to feel a sense of helplessness, to lack a sense of efficacy.

If there can be only, say, four A students in a section, and Ann is fifth in the group, then the only way she can earn an A is by "bumping" someone else. The "bumped" person need learn no less or become less competent, but in class curve-grading someone must receive a lower grade in order for Ann to receive a higher one. Thus in such systems, a student's grade is not only a function of the individual's achievement, but is also a function of the achievement of the others in the section.

Class-curve grading does not foster good interpersonal relations; rather, having to "bump" others and being "bumped" foster ill will. Grading on a class curve does not encourage group study or cooperative learning; instead, it encourages isolation and exclusion. Class-curve grading does not motivate students to help one another to learn; on the contrary, self-interest would be "best" served by interfering with the learning of one's fellows.

Criteria for Grading Systems

Some instructors seek to compromise the virtues and vices of the two major systems; e.g., use domain referencing for setting a passing standard and a class-curve system for differentiating among levels of passing, or use an average of a domain-referenced score and a norm-referenced score. Such approaches succeed in diluting the evils of each approach at the cost of simultaneously diluting the virtues of each.

In seeking better alternatives, we suggest striving to **avoid** (rather than dilute) the pitfalls identified for percentage and class-curve grading practices. We also seek to tap the undiluted values of each approach. These concerns give rise to the following bases for judging college grading systems.

Obtain Relevant Norm Referencing. Because the content domains of most college courses are large, flexible, and open and because instruction is aimed at transfer of learning rather than mere rote recall, it is impossible to describe the domain with sufficient precision either to render domain referencing meaningful or to make the establishment of content-mastery standards appropriate (cf. Hanna & Cashin, 1987). Therefore, the only viable means of interpreting performance is by means of norm referencing. A sensible system of college grading must be referenced to some **relevant** reference group of other people. The pursuit of relevance usually dictates that the reference group consists of other (past and/or present) students of the course.

Avoid Instability of Small Samples. Sound grading systems must also be referenced to **stable** groups. This requires groups large enough to avoid marked group-to-group fluctuation. The need for stability dictates that reference groups not consist of individual sections (unless sections are enormous). The grades awarded to students in a given section should be free to reflect the group's achievement if it turns out to be unusually high or low. Similarly, grades should reflect section-to-section oddities in dispersion.

Avoid Psychological Evils of Fixed-Sum Games. There is no preset total amount that the students in a section can achieve. Grading systems should reflect this reality. Student cooperation should not be thwarted by systems of reporting their achievement. Artificial competition among close peers should be avoided. This dictates that the reference group be (at least largely) **external** to the section being graded.

Provide Sense of Efficacy. Students should have a sense of control over their learning and over the grades that report their achievement. They should know that if the achievement is unusually high (or low) in their section, the grade distribution will reflect this status.

Be Defined and Interpretable. "One of the standards for any marking and reporting system is that it be as interpretable as possible" (Thorndike & Hagen, 1977, p. 590). A grading system's meaning should be communicable. The definition of grades should also be consistent from instructor to instructor and from section to section.

The Concept of Anchoring

Analysis of the two common grading systems led to an apparent paradox. Norm referencing was seen as the only logical foundation upon which to base college grades. Yet the only obvious means of achieving norm referencing was the unsatisfactory method of grading on a class curve. Anchor measures provide an escape from the apparent paradox.

An **anchor measure** is a device with which one can judge or "take bearings" of the status of a class. As an example, suppose several sections of the same college algebra course (i.e., sections all taught from the same syllabus, same text, etc.) all took the same final exam. This exam could be used to reveal if and how the groups differed in achievement, and the grade distributions in the several sections could be adjusted accordingly. Or suppose the prior SAT Math scores of students in each section were available. If some sections were found to have higher mean SAT Math scores than others, then the section-to-section grade distributions could be adjusted accordingly.

To provide anchorage, a variable need have only one attribute: it must correlate with performance in the course being graded. The greater the correlation, the better. Thus, common exams across sections would provide stronger anchoring than would prior GPAs. Notice that some anchor measures (e.g., common exams) can and should contribute to the evaluation of each student, while others (e.g., aptitude test scores or grades in prerequisite courses) clearly should not be used to evaluate student course work.

At their best, anchor measures help in attaining each criterion of sound grading systems. When a course achievement variable is used to anchor achievement of students in one section of a course to a large number of other students who have taken (or are taking) essentially the same course, the anchor measure provides the needed link to satisfy each criterion. First, the group of relevant other people who have had the same course provides meaningful, interpretable, relevant norm referencing. Second, the large size of the reference group provides stability from section-to-section sampling error. Finally, the ability of the individual section's grades to rise or fall with student achievement liberates students from a need to compete with peers and provides members of the section with a realistic sense of efficacy.

The statistical processes by which anchor measures are used can be relatively simple and intuitive. This is the approach used in the following examples to illustrate their use.

Examples

Instructor's Use of Final Exam as Anchor Measure

The first author teaches a small section of a graduate course each term. The final exam is revised only every few years (to coincide with text revision). This exam (or unrevised portions thereof) provides a measure with which the various sections can be statistically linked.

The syllabus announces that the distribution of grades in **typical or average** sections is about 30% A, 50% B, 18% C, and 2% D or F. It is explained that comparison is **not** with the small section, but with several previous sections of the same course. If a group's achievement is unusually high or low, its grades will reflect this; i.e., grading is **not** on a **class** curve.

In a certain section, there were only 10 students. Table 1 shows the cumulative distribution of final exam scores for the dozen previous sections and the distribution for the section under consideration. The percentage of previous students receiving each grade on the final exam approximated the announced distribution of grades in typical sections. The right-hand side of Table 1 reveals that the section under consideration exhibited higher-than-average achievement. This enables us to know the direction in which to adjust the grades that would have been given under conditions of class-curve grading—upward.

Now refer below to the distribution of **total points** accrued in the course (without digressing to how these points were assigned or summed). The question is, where to draw the lines separating the grades. There is no one correct answer. For the clearly superior class under consideration, one would want to award more than the typical 30% (or three) A grades. Given the relatively large breaks in the distribution between the fourth and fifth persons and between the sixth and seventh persons, one would likely award either four or six A grades. Likewise, given the gap between the bottom person and the others, one would likely assign only one C grade.

Distribution of Total Points

816
776
775
755
706
701
672
628
620
515

Table 1

Example of an Instructor's Use of Final Exam as Anchor Measure

Cumulative Distribution of Final Exam Scores for 12 Prior Sections		Final Exam Scores of Present Class	
Score	Frequency	Grade	Frequency
50			
49	2		
48	1		
47	3		
46	5		1
45	6		2
44	6	A	
43	6	29%	
42	10		
41	5		
40	9		2
39	12		1
38	17		1
37	9		
36	17		
35	8		
34	10		
33	9	B	
32	12	51%	1
31	10		
30	7		
29	5		
28	9		
27	11		
26	8		
25	4		1
24	2	C	
23	2	18%	1
22	5		
21	5		
20	3		
19	2		
18	1	D	
17	1	2%	
16	1		

All five criteria of sensible grading systems were easily met. The grades were rendered interpretable by being norm referenced to the larger, more stable, highly relevant group of previous students in the course. Yet the psychological and social evils of putting students in a fixed-sum pit were avoided. Finally, students were provided with a realistic sense of efficacy.

U. S. History in Community College

In the first example, the need to secure agreement among participating instructors was moot; there was only one. The next example illustrates the opposite extreme in which there is not enough similarity of content among the sections taught by different instructors to afford any common test.

Suppose three instructors each teach one or two sections of a United States history course. Although all are charged with teaching a general survey course, their methods and contents differ markedly. For example, two approach history chronologically, while one teaches it topically. They use different texts. Moreover, one tends to focus on political topics, one on economic and social history, and one on military topics. They cannot agree on common objectives or common exam content for the five 20-student sections.

In spite of these difficulties, several of the above criteria can still be met if the instructors can agree on one thing—the distribution of grades that would be suitable in **typical** sections; i.e., what fraction of students most often receive each grade? This consistent norm-referenced definition of the meaning of grades is an educational, social, and cultural definition, not a statistical or psychometric one (Thorndike & Hagen, 1977, p. 598). The instructors (or their department or institution) may have to engage in negotiation, compromise, arbitration, etc., to arrive at this definition. But it is worth the trouble; it **defines** grades!

Suppose they eventually agree that in **typical** sections, distributions should approximate 20% A, 25% B, 30% C, 20% D, and 5% F. (N.B.: We are not urging this or any other specific distribution or definition of grades. Rather, we urge local **consistency** for whatever distribution is agreed upon. We do, however, caution against a symmetrical array of grades on a normal curve because, as Frisbie, Diamond, and Ory, 1977, p. 12, said, "it is highly unlikely that college and university student achievement is normally distributed.")

This agreement on a definition of grades should have two components. First, each instructor should agree to abide by the decision and to give that approximate distribution in **typical** sections. When there is deviation from that distribution because a section is atypical, one should be willing to provide colleagues with credible evidence of the group's deviancy. Second, each instructor should not merely be authorized to deviate from a typical distribution when a group is atypical, but should be **charged** with doing so. If one were blindly to follow the typical distribution for each section, one would be grading on a class curve by referencing the grades to the one small section. Rather, the grades are to be referenced (via an anchor measure) to the larger, external group.

Clearly, the anchor measure will have to be something external to the class and therefore will not be suitable for student assessment. Suppose that the college requires all entering students to take the ACT. The Composite Score could provide the needed linkage. It would meet the fundamental condition of being correlated with course achievement.

Each section's distribution of ACT scores would be examined to see how it deviated from the distribution shape of the total group. This would then be used to adjust, where appropriate, the distribution of expected grades from what it would be in a typical section. Table 2 gives a sample of the kind of distribution of **expected** grades that the instructors might work out for the five sections in view of their prior ACT scores. No adjustment was needed in Section 1; its ACT scores were quite typical. In Section 2, ACT scores were a little below average; thus a modest downward adjustment was made for expected history grades. In Section 3, the adjustment was upward. The adjustments in Sections 4 and 5 are neither up nor down; rather they reflect differences in variability. Section 4 had unusually homogeneous ACT scores (i.e., center heavy with relatively few extreme scores); accordingly the expected distribution of history grades should be center heavy. In Section 5, the distribution is unusually heterogeneous.

Table 2

Expected Distribution of History Grades Based on Composite ACT Scores

In Typical Sections	In Section 1	In Section 2	In Section 3	In Section 4	In Section 5
A = 20%	A = 20%	A = 15%	A = 25%	A = 15%	A = 25%
B = 25%	B = 25%	B = 20%	B = 30%	B = 20%	B = 30%
C = 30%	C = 30%	C = 30%	C = 30%	C = 50%	C = 10%
D = 20%	D = 20%	D = 25%	D = 15%	D = 15%	D = 25%
F = 5%	F = 5%	F = 10%	F = 0%	F = 0%	F = 10%

Now suppose the semester is ending and the instructors have prepared the distributions of total scores for their sections shown in Table 3. The instructors' numeric scales will likely differ (like everything else about their teaching!), but the adjustments are still feasible. Use of anchor measures does not force instructors to adopt similar record keeping practices. In Section 1, the distribution of grades conformed to the expectations of a typical section except for the extra A grade in order to accommodate the tie. In Section 2, the natural breaks in the distribution enable perfect conformity with the expectations (listed in Table 2) based on the slightly below-average ACT performance of the class. The above-average distribution derived from the ACT scores of Section 3 were not quite as closely achieved in the distribution of final grades; this was because of the desire to make the grade divisions at the larger natural intervals or breaks in the distribution. The unusually center heavy distribution expected in Section 4 was approximated in the distribution of final grades, with the only departures being to accommodate natural breaks in the distribution of total scores. Likewise, the expectations for Section 5 were approximated in course grades.

How well were the criteria of good grading systems satisfied in this example? First, relevant norm referencing was achieved (albeit not quite as well as in the first example where the anchor measure was more relevant to the course content). Second, instability of small samples was avoided. Third, the psychological evils of a fixed-sum game were **not** avoided; once the anchorage has been achieved, the section became the new, adjusted reference group. Fourth, efficacy was **not** promoted; once anchorage had been effected, the target distributions were fixed. Finally, definition and meaning of grades was achieved. Owing to the instructors' inability to agree on any valid common achievement measures, only some of the benefits of anchorage could be realized.

Table 3

Total Point Distribution in Each Section

(Systems differ by instructor; 20 students in each section)

1		2		3		4		5	
345		349		99		98		1,107	
339		340	A (15%)	98	A (20%)	96	A (20%)	1,099	
337	A (25%)	338		96		95		1,090	A (30%)
336		330		96		95		1,081	
336		327	B (20%)	92		91		1,077	
325		321		92		90	B (15%)	1,076	
320	B (20%)	315		91		86		986	
315		299		91	B (40%)	85		980	B (20%)
309		290		91		85		960	
283		283	C (30%)	89		85		925	
280		276		88		85		871	
265	C (30%)	271		87		82	C (50%)	865	C (20%)
265		266		84		82		851	
264		259		82		80		843	
262		252		79	C (25%)	79		801	
249		252	D (25%)	79		78		737	D (20%)
241	D (20%)	239		76		74		732	
240		233		71		70		649	
238		199	F (10%)	68	D (15%)	68	D (15%)	510	F (10%)
105	F (5%)	170		65		64		305	

Multiple Sections Taught by GTAs in a Large University

Suppose a senior professor supervises and coordinates 20 graduate teaching assistants (GTAs) who each teach two small sections of freshman composition. With many inexperienced GTAs, there is a relatively greater need for anchorage in order to prevent wildly differing standards among the sections. Morever, in this situation involving relatively "junior," transient GTAs, there is a political potential for more supervision and articulation (even regimentation) than would be desirable for, or tolerated by, regular faculty.

Suppose the supervising professor has decided that 65% of student grades will be based on various writing assignments, 20% on an essay portion of a final exam, and 15% on an objective test of mechanics of written expression. The final exam could be common to all sections and thereby provide an excellent basis or anchorage. The only difficulty would be that GTAs and students alike would remain ignorant until the end of each term concerning how well a section and its students were doing vis-a-vis the reference group of all 40 sections.

A more viable approach would be to couple GTA training with securing essay anchorage early in the term. The GTAs could first be given practice and feedback in grading common sample materials. Next, all the GTAs could be asked to give students the same writing assignment. By having supervised evaluation of these written materials, each **by multiple GTAs evaluating work of students NOT in their own sections**, one could secure a good sense of the relative standing of the 40 sections. This would help each GTA to know how generously to scale letter grades for other written assignments (marked by only himself or herself). It would also provide a basis for anchoring the final distribution of grades in the respective sections. The objective test of mechanics of expression would provide a very convenient second basis by which to anchor the several sections' distribution of final grades.

Summary

Each of the common bases for assigning college grades—percentage and class-curve grading—has only limited meaning and lacks a sound rationale. Neither satisfies the major criteria for student grading systems. Anchor measures enable the advantages of norm-referenced grading to be achieved without introducing the evils of class-curve grading. Examples were used to illustrate how a variety of anchor measures can be used to achieve meaningful grading in varied contexts without intruding into instructors' record-keeping practices. They, therefore, are recommended for use in assigning college grades.

References and Further Reading

American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, DC: Author.

Dressel, P. (1983). Grades: One more tilt at the windmill. In A. W. Chickering (Ed.), *Bulletin.* Memphis: Memphis State University, Center for Study for Higher Education.

Frisbie, D. A., Diamond, N. A., & Ory, J. C. (1979). *Assigning course grades.* Urbana-Champaign: University of Illinois Office of Instructional Resources.

Hanna, G. S. (1985). *Coordinating instructional objectives, subject matter, tests, and score interpretation.* Manhattan, KS: Kansas State University.

Hanna, G. S., & Cashin, W. E. (1987). *Matching instructional objectives, subject matter, tests, and score interpretations.* (IDEA Paper No. 18). Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

McKeachie, W. J. (1986). *Teaching tips* (8th ed.). Lexington, MA: Heath.

Milton, O., Pollio, H. R., & Eison, J. A. (1986). *Making sense of college grades.* San Francisco: Jossey-Bass.

Terwilliger, J. S. (1971). *Assigning grades to students.* Glenview, IL: Scott, Foresman.

Thorndike, R. L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin.

Thorndike, R. L., & Hagen, E. P. (1977). *Measurement and evaluation in psychology and education* (4 th ed.). New York: Wiley.