*Bulletin of the Psychonomic Society*
*1983, 21 (3), 209-212*

# Grade inflation as a legitimate response to the unreliability of teacher-made tests for university-level coursework

ALLEN J. SCHUH
*California State University, Hayward, California*

**Major objections to the A-to-F grading scheme include grade inflation and unreliability of the teacher-made test. The grading model proposed here recommends the use of conventional objective-item tests and calculates an estimate of the test's reliability. The reliability estimate is then used in the calculation of the standard error of measurement. It is suggested that a two-standard-errors-of-measurement cushion undercutting scores for the next higher grade level should provide ample latitude in the assignment of grades and that this form of grade inflation is justifiable. There is little that is truly original in this article, since the formulas have been in the literature for over 40 years. If there is a contribution, it is in the presentation of arguments on the issue of grade inflation and in the attempt to show how the time is now ripe to apply these well-known procedures to justify one form of grade inflation.**

One needs to give careful consideration to the assignment of grades for student coursework at the university level. Grades are more important today than ever before for entrance into graduate school, successful employment, tuition refunds, and parental respect. At the same time, some people perceive grades to be more important than the knowledge the grades are supposed to signify. Grades have been criticized on many grounds, such as that grades: (1) maintain a racist, elitist, and sexist educational system; (2) foster destructive competition; (3) create a Watergate morality; (4) divide students and teachers; (5) reduced learning to a survival-of-the-fittest charade; (6) devalue self-worth; (7) lack reliability; (8) keep teachers in line; and (9) inflate themselves (Bellanca, 1977; Kirschenbaum, Napier, & Simon, 1971; Pinkus & West, 1980). Despite the variety or even the validity of charges against the grade system, no one seriously anticipates that the A-to-F grading system will disappear in the forseeable future. For example, at the university at which the author teaches, the catalog mandates that A, B, C, D, F grades will be assigned in all coursework required for the major in the major department. Perhaps the efforts of professors should begin to focus more seriously on assigning the traditional grades according to methods that are more reliable and more standardized for the professor, the coursework, and the population.

One should recall that, in grading, the professor records a written evaluation of the performance by a student in designated coursework by assigning a letter grade of A, B, C, D, or F. The grade A denotes the highest level of accomplishment. B represents higher accomplishment than C. The grade D or F denotes performance at less-than-acceptable levels. If a student were to achieve all Cs but one, and that grade were either a D or an F, the student would be placed on academic probation at some institutions. Such probation could lead to withdrawal from the university without a degree, because a student on probation cannot graduate. Thus, the assignment of a letter grade of D or F is a threat to academic standing. The author believes that avoiding erroneously low grades is a better strategy than avoiding erroneously high grades.

Whereas a grade of D or F can threaten a student's quest for the baccalaureate, even a grade of B can threaten the student's future career interests if the student intends to pursue a professional degree in law, medicine, engineering, or the like at one of the prestigious universities. There are just not enough openings in prestigious professional schools to accommodate all of the qualified A students. Thus, a B student is simply no longer in the zone of consideration.

It is possible, then, that some grade inflation occurs because professors do not believe that they are justified in blocking a student's present pursuit of the baccalaureate or future career interests. One might expect this to be true especially when a student falls just below a cutoff score for the next higher grade and the professor legitimately doubts whether the grading scheme is sufficiently reliable and valid to make the rigid establishment and enforcement of cutting scores justifiable to oneself, the administration, or the student.

Everything considered, the use of undergraduate grades for graduate-school screening is probably justified, because reading level, verbal understanding, motivation to do well, and effective work habits tend to put

The author wishes to thank Howard Rosenberg, William L. Sawrey, and John M. Stevens for their comments on the earlier version of this paper. The author's mailing address is: Department of Management Sciences, California State University, Hayward, California 94542.

 

the same students ahead in most learning situations. Put another way, a substantial body of evidence shows that the single best predictor of future behavior is past behavior in a similar academic situation (Cronbach, 1970). The external validity of academic grades is questionable. Since school grades can be very poor predictors of real-world achievement, leniency helps prevent exclusion of later achievers from graduate education opportunities.

Of course, two phenomena that may have produced another kind of grade inflation have occurred in the past decade. On the one hand, choice academic positions are more scarce because of declining student enrollments. On the other hand, undergraduate students are now frequently allowed to record their opinions of the professor's teaching effectiveness. Thus, a certain amount of grade inflation might be simply a personally rational response by the professor to a reward system that pays off for giving high grades. Student assessments of a professor's teaching effectiveness are a major source of information in administrative promotion and tenure decisions. And, students reciprocate to the professor the evaluations they themselves have received (Worthington & Wong, 1979). Thus, "good" professors assign generously high grades to their students. These "good" students reciprocate with generously high grades to the "good" professors, and one is quickly into a deviation-amplifying loop.

Even if one were to hold constant the student evaluations and the leniency of some professors in grade assignment because of keener competition for grades by some students, it is possible that some grade inflation would occur following the use of more effective instructional strategies such as mastery learning (Denton & Henson, 1979). Of course, under those circumstances, there would be, technically, no real "inflation," because the grade still would represent what it used to represent.

How is one to judge, given knowledge that a particular student has received straight As, whether the student really is that good? Or, are the grades the result of systematic errors: lenient professors who compensate for the unreliability of the system, professors desperate for tenure who generously reward students so that the students will generously reciprocate on the student evaluations of their teaching effectiveness, or the high level of productivity made possible through better structuring of courses and course materials and innovative strategies of instruction? A detached observer, of course, has no way of knowing which, if any, of these singly or in combination are the reasons for a particular student's getting A grades.

If we are going to perpetuate the A-to-F grading scheme, and it appears we shall, then there is a clear need for an assignment model that preserves as much as possible a reasonable and justifiable range of grades. The model should be robust in being standardizable across professors, across courses, and perhaps even beyond one geographical area. The model should make good psychometric sense and be easy to calculate and to understand.

It would be desirable to have it appear that a systematic grade-awarding scheme, and not the professor's subjective evaluation of performance, determines what grade the student gets. This would help to neutralize the reciprocation error (Worthington & Wong, 1979). Thus, the professor could spend the available time developing course and examination content for presentation in a structured format. Such procedures would mitigate the perceived power role of the professor (Schuh, 1978).

It is, therefore, to the professor's advantage to teach the material to the best of one's competence and to attempt to increase test reliability, and then to let the students' grades fall where they may along a rationally preestablished framework. If estimates of test reliability are available for the professor's homemade tests, that information itself might be useful to promotion and tenure committees. Perhaps the reliability information might be even more meaningful than the students' opinions of the professor. Professors who have reliable examinations and grade generously do so more credibly than professors with tests of low reliability. Grade information is also used by the students, and grade feedback can affect the students' subsequent motivation to learn and their self-concepts.

## PROCEDURES WITH A THEORETICAL EXAMPLE

The professor should break the course material into several units of equal size and difficulty. Objective examination items should be developed and presented in accordance with normal testing procedures. The grading scheme proposed here tells the professor what to do with the resulting test performance to make the grade-awarding scheme more defensible.

All of the steps proposed here can be done with the use of pocket calculators. The formulas suggested are generally available with a thorough discussion in textbooks on educational and psychological testing (Anastasi, 1976; Cronbach, 1970; Guilford, 1954; Gulliksen, 1950; Nunnally, 1970).

One of the more concise guides to teacher-made test construction (Educational Testing Service, 1959) suggests that the professor can construct objective items (multiple choice, matching, completion or fill in, and true-false) for a test that can measure the students' knowledge just as effectively as can essay examinations. A large number of items should probably be developed to test in each major category of the cognitive domain: knowledge, comprehension, application, analysis, synthesis, and evaluation (Bloom, 1956). Enough items should be presented such that all relevant areas are covered and about 90% of the students can complete all of the questions on the tests. The larger the number of items that sample effectively the material in lectures and tests, the greater will be the reliability of the evaluation procedure. Each item on the test should have the usual psychometric properties of variance (some, but not all, of the students get the item correct) and item-total test correlation (those who score high on the total tests are more likely to get each item correct). To avoid ceiling

effects, the test should have at least some items hard enough so that no one gets 100% of the items correct.

(1) Consider a test with 100 objective items and in which each item is scored as correct or wrong. One knows that the highest score will receive a grade of A. For example, if the highest grade is 97, that is, the best student missed 3, then that is by definition the highest A.

(2) One should examine the theoretical chance distribution to judge which students did significantly better. Calculate the theoretical mean and standard deviation for the distribution of scores that would result if everyone answered the test in a random manner. Thus, for a 100-item true-false test, a mean of 50 would be expected by chance, since the mean of a binomial test is np, or the number of items in the test times the probability of passing each item. The standard deviation of this distribution is $\sqrt{npq}$, where $q = 1 - p$. For example, $\sqrt{npq}$ with a 100-item true-false test would be 5.

If one expects students who score significantly better than chance to be awarded a letter grade of C, then it follows that the professor should calculate 1.65 standard deviation units above the theoretical mean, which is the one-tailed value at the 5% level of statistical significance (in this example, 8.25). Thus, any student who misses only 41 items is doing at least better than chance. Fifty items plus 8.25, rounded off to the next whole number, is 59 right, or 41 wrong.

(3) If one knows the highest A (Step 1) and lowest C (Step 2), one might simply divide the distance between them by 3, and B territory then falls to the middle ground. These are tentative cutoff scores only, and they require further refinement. With the highest score of 97 (or −3) and lowest allowable C of 59 (or −41), there are 39 points between 97 and 59; thus, 13 points are assigned to As, Bs, and Cs. The grade range is then 97 to 85 for A, 84 to 72 for B, and 71 to 59 for C.

(4) The role of chance in test performance and especially in students' scoring just below cutting scores must be explicitly acknowledged and used to protect students who just miss the next highest grade through sampling error. If one protects the downside risk of error, there is apt to be less criticism of the procedure as a whole.

A common procedure in test performance interpretation for aptitude and achievement tests is to report the standard error of measurement (Dudek, 1979) and to acknowledge that a person's true performance on a single occasion is probably within plus or minus two standard errors of obtained performance. It is suggested that such an interpretation of student performance be used when test performance falls just short of the cutting score for the next higher grade.

Put another way, some grade inflation is justified. Students should not be penalized for the lack of reliability in the teacher-made test. The high importance of grades requires the professor to give the student the benefit of the doubt when the student's performance is within two standard errors of the next higher grade, especially if that next higher grade is an A or a C, because of the personal and administrative consequences of error here. A fallout of this procedure is that the number of students who fall within each grade zone is less important than whether the grade zone itself is justifiable on psychometric grounds. Thus, all the students in a class could get an A on a test or in a course, and it could be defensible to an audience with credentials in psychometric evaluation. It might be assumed that giving all As (or all Fs) would embarrass the academic administration, but only if the grade-awarding scheme were without a quantitative foundation.

The lowest numerical score to be awarded the next higher grade level should be that score that falls two standard errors of measurement below the tentative cutoff scores in Step 3 above. The standard error calculation requires knowledge of the test's reliability. The test reliability can be calculated with KR21 (Kuder-Richardson Formula 21):

$$r_{tt} = \frac{n}{n-1}\,\frac{\sigma_t^2 - n\bar{p}\bar{q}}{\sigma_t^2},$$

where $r_{tt}$ is the estimated reliability, n = the number of items in the total test, $\sigma_t^2$ = the variance of the total test, $\bar{p}$ = the mean of the test divided by the number of items, and $\bar{q} = 1-p$.

Thus, this formula requires only the mean and standard deviation of the distribution of test scores and the number of items on the test.

One sets the actual cutoff score for the test, or the total accomplishment in the course, at two standard errors of measurement below the tentative cutoff score.

The major determinants of how far below the tentative cutoff scores one sets the final cutoff scores depends on the standard deviation of the distribution of student scores and on the reliability of the teacher-made test. Thus, if the standard deviation of student performance on this example test is 7.00, the reliability of the test will determine how much further one drops down to ensure, within two standard errors of measurement, that one is awarding enough grades of the next higher rank. For reliabilities of .90, .70, .50, and .30, the additional latitude required would be 4.34, 7.56, 9.80, and 11.62. In each case one simply applies the formula $SEM = S\sqrt{1 - r_{tt}}$, where SEM = standard error of measurement, S = standard deviation of student scores on the test, and $r_{tt}$ = the estimated reliability of the test based on KR21. Thus, although the tentative cutoff score for an A was 85, if the reliability is only .30 one should drop down to 85 − 11.62 and round to a whole number, or 73, to include those who are really As but who scored low because of sampling error and the unreliability of the test.

## CONCLUSIONS

Professors have to assign grades to students for coursework even though they (professors) may be very familiar with some of the negative factors that are associated with the whole grading

system. Whatever grading scheme a professor adopts should be robust in allowing comparisons over time between academic terms for a course, between courses for a professor, and even between professors for a course. Such a robust grading methodology should be applicable to a wide variety of courses in a university curriculum. It then follows that professional schools are more apt to trust the grade reports from a university program in which sound psychometric procedures have been used in awarding grades than from one in which grading procedures are somewhat less standardized and, therefore, are more in doubt.

This paper presents the author's perspective and a solution that answers at least some of the criticisms on how grades are frequently awarded. The proposed solution confronts the test reliability problem directly by calculating its level and using it in turn to determine the standard error of measurement. The two-standard-errors-of-measurement cushion that a professor can build in below tentative cutoff scores is at least a partial assurance that students are not penalized for the less-than-perfect reliability of the teacher-made test. Other ways to reduce error variance (i.e., via longer tests, more frequent tests through the term, or use of pretested items) seem to the author to be less practical than the steps suggested here. It is hoped that other professors in other places will attempt to apply the method suggested here. Wiser and more satisfactory awarding of grades may result.

## REFERENCES


ANASTASI, A. *Psychological testing*. New York: Macmillan, 1976.

BELLANCA, J. A. *Grading*. Washington, D.C: National Education Association, 1977.

BLOOM, B. S. (Ed.). *Taxonomy of educational objectives* (Vol. 1): *Cognitive domain*. New York: McKay, 1956.

CRONBACH, L. J. *Essentials of psychological testing*. New York: Harper & Row, 1970.

DENTON, J. J., & HENSON, K. T. Mastery learning and grade inflation. *Educational Leader*, 1979, **37**, 150-152.

DUDEK, F. J. The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin*, 1979, **86**, 335-337.

EDUCATIONAL TESTING SERVICE. *Making the classroom test: A guide for teachers* (Evaluation and Advisory Service Series No. 4). Princeton: Author, 1959.

GUILFORD, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.

GULLIKSEN, H. *Theory of mental tests*. New York: Wiley, 1950.

KIRSCHENBAUM, H., NAPIER, R., & SIMON, S. B. *WAD-JA-GET? The grading game in American education*. New York: Hart, 1971.

NUNNALLY, J. C. *Introduction to psychological measurement*. New York: McGraw-Hill, 1970.

PINKUS, A. G., & WEST, J. V. An alternative method of grading to letter grades and percent scores: "Relative order." *Journal of Chemical Education*, 1980, **57**, 89-90.

SCHUH, A. J. Variations in lecture task orientation and student perceptions of course effectiveness. *Bulletin of the Psychonomic Society*, 1978, **11**, 193-194.

WORTHINGTON, A. G., & WONG, P. T. P. Effects of earned and assigned grades on student evaluations of an instructor. *Journal of Educational Psychology*, 1979, **71**, 764-775.


(Received for publication March 5, 1983.)