Reprinted from PSYCHOLOGY IN THE SCHOOLS, Vol. II, No. 1, 17-24, January 1965

GRADES AS MOTIVANTS[1]

LEWIS R. GOLDBERG

University of Oregon, and
Oregon Research Institute

There is probably no aspect of higher education so important to students and so exasperating to educators as grading, yet it is easily one of the least-studied phenomena of the college scene (Lamson, 1940; Odell, 1950). There has been long-standing interest in the characteristics of achievement tests and methods of arriving at individual scores, but marked neglect of the effects upon later learning of various methods of converting a distribution of scores to a distribution of letter grades. While there are many ways of reporting student achievement other than letter grades (e.g., rank in class, percentiles, etc.), letter grades are probably the most common academic reporting system, and in most American institutions of higher learning even other reporting systems must ultimately be converted into letter grades by the term's end.

The recurring problem of how best to convert test scores to grades may present as much anxiety to the instructor as does the anticipation of poor grades to the student. There are at least three types of pressures (not counting the instructor's own humanitarian impulses) which impinge on his grading decisions. The first of these is *administrative*. The institution, and increasingly the wider circle of institutions of which any one college is only a part, press for comparability in grading procedures across different courses, different instructors, and different academic years (e.g., Anderhalter, 1962). The administrative demand is for comparability in the meaning of grades, the polar extreme of which is typified by programs utilizing objective tests, administered by persons outside any one academic institution, with a pre-set and *invariant* procedure for reporting test performance (e.g., the College Board examinations, the Graduate Record examinations, etc.).

A second pressure on the instructor in converting test scores to letter grades is an *informative* one. Since a major rationale for grading is to provide information to the student on his progress in the course, this informative pressure pushes the instructor to adopt a conversion system which provides the maximum amount of feedback possible, while at the same time constraining him from making distinctions finer than can be reliably made from the particular test administered.

A third source of pressure on instructors in arriving at grades—especially grades on quizzes, mid-term examinations, and other tests given earlier than the very end of the course—is a *motivational* one. Typically, the instructor may wish to choose a grading distribution which will maximally motivate his students to learn more during later parts of the course. It is this third aspect of grading (which provides some supposed leverage by the instructor on future student learning) that engenders the most heated debate among academicians, since each instructor has his

[1]The author wishes to acknowledge the help of William Resch and Bruce Palmer who assisted the project both as instructors and as project assistants, of Herb Penney and Lloyd Lovell who provided encouragement and support, and the entire staff of Oregon Research Institute who provided stimulation and critical reading of an earlier draft of this paper.

Funds were provided by the Oregon State System of Higher Education and the cost of data analysis was defrayed by the Computing Center of the University of Oregon. Thanks are due to Kuo-Cheng Hsieh, Robert Gregovich, Terry Liittschwager and Linda Dutcher who assisted with the statistical analyses.

own intuitive position regarding the grade distribution most conducive to increasing later performance.

One such intuitive model of motivation, not uncommonly held by academicians, can be viewed as a "lenient" grading policy. The motivational model of the lenient grader may rest on the psychological assumption that pleasure is a more effective motivant than pain, and that optimum learning occurs under conditions of frequent positive reinforcement (coupled with infrequent errors and consequently infrequent non-reinforcement and/or punishment). The lenient grader—that is, the instructor who converts his test score distributions to letter grades such that there is a markedly skewed distribution of high grades—frequently assumes that (a) high grades are positive reinforcements, which in turn (b) affect learning *directly* through an increase in the probability of making a subsequent response similar to that reinforced, and (c) *indirectly* through the effect of generalization of pleasure to *learning as a process.*

The polar opposite of the lenient grader could be viewed as the "strict" grader —who gives, at least on early examinations, mostly *Ds* and *Fs* and uses *As* and *Bs* sparingly if at all. He typically assumes (a) that most students are working far below the level at which they are capable, and (b) that punishment, in the form of poor grades, is the best motivational inducement to encourage more intensive efforts at learning. While these models are probably overdrawn for any one college instructor, they may illustrate the range of belief regarding the motivational effect of grades—from beliefs analogous to those expressed in *Summerhill* (Neill, 1960) on one extreme to the philosophy of a Marine Corps boot camp on the other.

A third motivational model, undoubtedly less commonly held than either the strict or lenient models, stems from expectancy theories of motivation, of which the best example can be seen in the early writings of Hebb and, later, in the theoretical positions of the need-Achievement investigators (e.g., McClelland, Atkinson, Clark, and Lowell, 1953). Instructors of this stance would argue that neither positive nor negative reinforcement, in and of itself, provides the best motivant, but rather the *difference* between what the student expects and what he actually receives provides the motivational punch. An *extreme* grade, being most divergent from the expectations of most students, would thus be the most motivating, regardless of whether it was a high or a low one. Proponets of such a grading rationale might tend to grade *bimodally* on early examinations, giving relatively many *A*s and *F*s and comparatively few, if any, *C*s. Since grades for most students would differ from their expectations, the poorer students might be encouraged to work harder to get better grades while better students might be encouraged to work harder to keep their high grades.

These three motivational models—leading to great diversities of grading conversion policies—can probably be found to be reflected, to a greater or lesser extent, in the beliefs of instructors in any institution of higher learning. For psychologists, interested in motivational models in their own right and faced with grading policy decisions of an applied sort, the opportunity to use this common college problem for naturalistic experimental research is a challenge. The present study represents an attempt to examine the motivational effect of the three previously mentioned grading policies, and two control policies, upon later performance in a college course.

Procedure

General

Five different grading policies were utilized in reporting the results for a first mid-term examination in a large General Psychology course and the comparative effects of these policies were tested by examining performance on a second mid-term examination, administered about one month later. Since the students in the course had previously been divided among four instructors, each utilizing a different textbook, this research can be viewed as four concurrent studies of the effect on subsequent learning of five different grading policies.

The Course: Experimental Subjects

The experiment was run during the second quarter of a three-quarter sequence in General Psychology at the University of Oregon. All students in the course were asked to view three televised lectures per week. In addition, each student was assigned to a one-hour-a-week discussion section, taught by one of four graduate assistants in the Psychology Department. Each of these instructors used a different textbook. The students read approximately one-half of the textbook (having covered the other half of the book during the preceding quarter) plus 28 reprints from the *Scientific American* (having read 20 different reprints during the preceding quarter). Each instructor taught five discussion sections.

Six hundred and nineteen students took the first mid-term examination, administered about three weeks after the beginning of the quarter. These students had been divided on a non-systematic basis in approximately equal numbers among the four instructors and, for each instructor, among his five discussion sections. With the usual attrition due to early drop-outs and change of sections due to students' scheduling conflicts, the number of students in a discussion section ranged from 22 to 44. The first mid-term examinations for each instructor's five discussion sections were graded using a different one of five different grading policies. That is, each instructor taught five sections, each of which was graded under a different policy. Approximately four weeks after the first mid-term examination, a second mid-term examination was administered; approximately three weeks later, at the end of the quarter, the final examination in the course was administered. Students were told that the first mid-term examination would contribute only 10 per cent to their grade in the course, that the second mid-term examination would contribute 40 per cent, and that the final examination would make up the remaining 50 per cent of their course grade.

Letter grades were reported back to each student after each mid-term examination during the discussion section following the examination. After the second mid-term examination, students were told of the experimental procedures used in grading the first mid-term and were told that a uniform grading policy would subsequently be applied to their first mid-term examination scores.

All three examinations were multiple-choice in format, filled out on IBM answer sheets, and machine-scored using the IBM 805 scorer. Each examination was composed of items covering the televised lectures (approximately 35 per cent of the questions), the textbook (approximately 45 per cent of the questions), and the *Scientific American* reprints (approximately 20 per cent of the questions). The textbook questions were culled from the instructor's manual for each textbook (rewritten, if necessary, for greater clarity and/or relevance); lecture and *Scientific American* items were specially written by the four instructors plus an additional, hired, test-item writer. There were 50 items on mid-term examination I, and 75 items on mid-term examination II.

The Five Grading Policies

In order to ascertain most fully the motivational effect of each grading policy, grading curves were made relatively extreme. Thus, while the actual grading policies utilized in this study may be more extreme than those typically employed in most college courses, they should allow for the full emergence of their motivational impact.

Strict, lenient, and bimodal grading policies were compared with two types of "control" policies: normal curve grading and rectangular grading. Table 1 lists the percentage of students receiving each letter grade for each of the five grading policies.

Table 1. Percentage of Students Assigned Each Letter Grade for Five Grading Policies

Grading Policy	Percentage of students who received *A*	*B*	*C*	*D*	*F*
Strict	.00	.05	.25	.40	.30
Lenient	.30	.40	.25	.05	.00
Bimodal	.20	.30	.00	.30	.20
Normal	.10	.20	.40	.20	.10
Rectangular	.20	.20	.20	.20	.20

Score distributions for the total mid-term I test were computed for all students taught by an individual instructor; that is, four score distributions were computed, one for each instructor, pooling

the test scores of all students in his five sections. Each of the instructor's five discussion sections was randomly assigned a particular grading policy. Letter grades for each student in each discussion section were arrived at by simply cutting the total distribution so as to yield the required percentages of letter grades for that section.

It is important to note that grades, themselves, were *not* assigned randomly (a grading policy which may characterize students' perceptions of some instructors' grading but which is rarely, if ever, explicitly espoused by diligent instructors). A student who achieved a score in the top five per cent of those taking the test would receive a *B* under the strict policy, and an *A* under all other policies. Conversely, a student who achieved a test score in the bottom five per cent would receive an *F* under all policies except the lenient one, where he would receive a *D*. A student falling near the middle of the test score distribution would receive a *C* under normal and rectangular policies, a *B* or *D* under the bimodal policy, a *B* under the lenient policy, and a *D* under the strict policy. For the first mid-term examination, only letter grades were reported, and requests for individual test scores were denied. However, all students received information on the correct answers to all test questions, prior to receiving their letter grades, and those students who remembered their own responses could gain a rough idea of their actual test scores.

Results were analyzed by analysis of covariance (Winer, 1962, pp. 578-605), programmed for the IBM 1620 computer, using the method of unweighted means (i.e., utilizing the harmonic mean) to adjust for unequal cell sizes.

Results

Analysis Over All Five Grading Policies

Analysis of Mid-Term II Total Scores. If subjects had been *randomly* assigned to their discussion sections, an analysis of variance for mid-term I scores should have indicated no significant differences between sections, and consequently a simple analysis of variance on mid-term II scores would be appropriate to reveal any motivational effects arising from the five grading policies. However, in any assignment system where the student can exercise even some limited choice in his assignment (e.g., time of day), systematic biases may occur. And, in fact, initial analyses of variance for mid-term I scores indicated some significant differences between sections on the mid-term I examination, thus making the initial analyses of mid-term II scores ambiguous. Consequently, it was necessary to run analyses of covariance on mid-term II scores, covarying out the scores earned in the mid-term I examination. For an extended discussion of the merits of this form of statistical analysis in educational research, see Fitch, Drucker, & Norton (1951).

Four such analyses were carried out, one for each of the four insturctors, since each instructor used a different textbook and consequently part of the examination was different for each instructor. Thus the results can be viewed as four replications using different instructors and textbooks of the effects of five different grading policies. All four analyses of homogeneity of within-class regression coefficients revealed no significant differences. All *F*-tests were not significant. Consequently, it would appear that the experimental grading policies had no significant differential effect upon later learning for any of these four instructors' classes.[2]

Analysis of Mid-Term II Part Scores. Since both the mid-term I and mid-term II examinations contained two sections which were identical for all four instructors (questions covering the televised lecture material and questions covering the *Scientific American* readings), a two-way analysis of covariance was possible utilizing the common portions of all examinations. This analysis would indicate if there were

[2]Additional tabular material has been deposited as Document number 7880 with the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C. A copy may be secured by citing the Document number and by remitting $1.25 for photoprints, or $1.25 for 35 mm. microfilm. Advance payment is required. Make checks or money orders payable to: Chief, Photoduplication Service, Library of Congress.

any interaction effects between instructors and grading policies. Again, analysis of homogeneity of within-class regression coefficients proved non-significant. In addition, no significant differences resulting from the five grading policies, the four different instructors, or the interaction between them emerged. From the results of both analyses, then, one might conclude that grading policies—even as extreme as those utilized in this study—have *no* differential motivational impact.

However, perhaps different grading policies may have had differential effects for different types of students. Perhaps high achievement students profited from one type of curve while low achievement students were more motivated by another. To test this hypothesis, the sample was divided into high, medium, and low thirds on the basis of their previous college grade point averages, and analyses of covariance carried out. These provided a two-way analysis of covariance (mid-term I scores covaried out) testing the effect of (a) grading policy, (b) cumulative college grade point average, and (c) any possible interaction between G. P. A. and grading policy upon mid-term II scores. While college G. P. A. was highly related to mid-term II scores (even with mid-term I scores covaried out), no significant interaction effect was uncovered.

Analysis between Strict and Lenient Grading Policies

The fact that no significant grading policy effects were uncovered may arise as a consequence of the sheer number of different grading policies utilized in this study (since as the number of intermediate experimental treatments increases, the probability of finding a significant difference between two extreme treatments decreases; Dixon & Massey, 1957, pp. 256-257). For this reason, a subsequent series of analyses was carried out for the two most extreme grading policies, the strict curve vs. the lenient curve (see Table 1). These curves probably encompass the range of grading procedures now used, and each has at least some advocates among college instructors. While these analyses are admittedly *post hoc*, they provide data relevant to the question: "If one is faced with a decision whether to grade strictly or leniently on an initial examination, which type of grading policy will most affect performance on a later examination?"

Four analyses of covariance were performed on mid-term II total scores, with total scores on the mid-term I examination covaried out. All tests of homogeneity of within-class regression coefficients proved non-significant. While three of the four strictly graded groups achieved higher scores than did their leniently graded counterparts (averaging about three items higher on a 75-item test which yielded mean scores of about 50), only one analysis of covariance reached statistical significance.

The findings for mid-term II part scores (mid-term I part scores covaried out) for strict vs. lenient grading policies and for all four instructors again showed that strictly graded students achieved slightly higher grades on mid-term II than did the leniently graded students, and the analysis of covariance indicated that this difference was statistically significant. No significant instructor effects emerged, and no significant interaction occurred.

The sample was divided into thirds on the basis of cumulative G. P. A., and an analysis carried out on strict vs. lenient policies and high vs. medium vs. low G. P. A. students. Both grading policy and G. P. A. affected mid-term II scores ($p < .05$), but again no significant interaction effects emerged.

In summary, then, a series of analyses of covariance showed no significant differences between five grading policies in their effect upon later performance on an examination. If one views solely the most extreme curves (strict vs. lenient), a smidgeon of *post hoc* evidence tended to favor the strict curve over the lenient curve, but the magnitude of the effect was very small and the number of analyses carried out makes interpretation of these significance levels risky. However, if one wished to make a case for strict grading on the basis of these findings, then one should probably also consider *other* differential effects of various grading policies, such as effects on student satisfaction with the course, likelihood of taking other related courses, etc. One such ancillary effect is the percentage of students dropping the course (when this option is allowed) as a function of various grading policies. Table 2 presents the drop-out percentages for the five grading policies and the four instructors.

TABLE 2. PERCENTAGE OF STUDENTS DROPPING COURSE AS A FUNCTION OF GRADING POLICY AND INSTRUCTOR

	N	Percentage Dropping Course
Lenient	112	.01
Strict	128	.08
Bimodal	119	.04
Normal	125	.02
Rectangular	135	.05
Instructor-Text A	165	.04
Instructor-Text B	154	.02
Instructor-Text C	144	.07
Instructor-Text D	156	.03
Total	619	.04

Since the last day on which University regulations permitted students to drop a course occurred between the first and second mid-term examinations (i.e., after the introduction of the experimental grading procedures), this analysis allows a clear-cut evaluation of the effects of (a) four different instructors and (b) five different grading policies upon course drop-out rate. As Table 2 indicates, the strictly graded sections had more drop-outs than did the leniently graded sections, and a Chi-square test of the significance of the differences between all five grading policies was significant ($\chi^2 = 9.94; df = 4; p < .05$). Differences between instructors were not statistically significant. Consequently, it would appear that one effect of strict grading is a slightly increased course drop-out rate and this effect might be considered, along with the small apparent differences in later test performance, in arriving at a choice between grading policies.

DISCUSSION

The findings from this study—that five different grading policies had little differential effect on subsequent test performance—should force proponents of a particular grading policy to make a thorough reappraisal of their beliefs. And, while instructors may still wish to defend their own grading practices on *admini-*

strative and/or *informational* grounds, they should now be wary of including *motivational* rationales as part of their arguments.

Moreover, the only published study which the author was able to find that has even tangential relevance to the present research provides no rebuttal to these findings. Billingslea and Bloom (1950) divided 24 seniors in an advanced psychology course into two matched groups, one of which was arbitrarily given failing grades on two essay examinations, the other given passing grades. The investigators expected to find a marked decrease in lecture note-taking on the part of the "failed" group as compared to the "passed" group, but analysis of the notebooks of both groups of students failed to find statistically significant differences.

Nor can the present findings be argued away solely on the basis of the types of grading curves utilized in this study. Kirby (1962) computed the average grades given by 206 instructors of lower division courses in *one* university and found a range from 1.9 to 3.9 (on a 4-point scale). It is apparent that the grading curves utilized in the present study (which would have yielded means ranging from approximately 1.0 to 3.0) were in general more strict than those representative of university courses (since the present curves represent grading policy deviations from a normal curve, while more typically instructors deviate from a skewed lenient grading curve). Therefore, the most "typical" comparison from the present study might well be the lenient curve vs. the normal curve. Here again, however, *no* significant differences on later test performance were found.

Finally, it is important to bear in mind that even the slight differences in test scores noted between the strict and lenient curves could have occurred primarily as a consequence of the differential drop-out rate between the two grading policies. While it is likely that primarily lower ability students dropped out of the strictly graded sections (and thus the average ability level of the strictly graded sections was slightly higher than that of the leniently graded sections at the time of the second mid-term examination), the *covariance* analysis utilized in the present study (which adjusts for differences between samples in their proportion of lower ability students, insofar as this is reflected in actual scores on the first mid-term examination) rules out the possibility that this ability level difference itself produced the slight differences in mid-term II test scores. However, it is not unreasonable to assume that among the lower ability students there are individual differences in studiousness, and that it is the less studious students who will more likely drop a course, given an initial poor grade. Since the strictly graded sections—with their increased drop-out rate—will contain fewer such students than their leniently graded counterparts, the slight apparent differences between these two groups may reflect merely the resulting disproportion in the distributions of these students. If this hypothesis is correct, a replication of this study in a school where dropping a course was not allowed should yield *no* differences, even between the extreme strict and lenient policies.

To summarize, then, while it is not argued that specific grades can not be used as "rewards" and "punishments" to elicit or inhibit specific attitudes or behaviors (e. g., Bostrom, Vlandis, & Rosenbaum, 1961; Ring & Kelley, 1963), it would appear that grading *policies* have negligible differential motivational effect on subsequent test performance at the college level.

References

ANDERHALTER, O. F. Developing uniform grading standards in a university. *Journal of Experimental Education*, 1962, *31*, 210-211.

BILLINGSLEA, F. Y., & BLOOM, H. The comparative effect of frustration and success on goal-directed behavior in the classroom. *Journal of Abnormal and Social Psychology*, 1950, *45*, 510-515.

BOSTROM, R. M., VLANDIS, J. W., & ROSENBAUM, M. E. Grades as reinforcing contingencies and attitude change. *Journal of Educational Psychology*, 1961, *52*, 112-115.

DIXON, W. J., & MASSEY, F. J., JR. *Introduction to statistical analysis.* (2nd ed.) New York: McGraw-Hill, 1957.

FITCH, M. L., DRUCKER, A. J., & NORTON, J. A., JR. Frequent testing as a motivating factor in large lecture classes. *Journal of Educational Psychology*, 1951, *42*, 1-20.

KIRBY, B. C. Three error sources in college grading. *Journal of Experimental Education*, 1962, *31*, 212-218.

LAMSON, E. E. The problem of adequate evaluation of the college student's achievement. *Education Administration and Supervision*, 1940, *26*, 493-507.

MCCLELLAND, D. C., ATKINSON, J. W., CLARK, R. A., & LOWELL, E. L. *The achievement motive.* New York: Appleton-Century, 1953.

NEILL, A. S. *Summerhill; a radical approach to childrearing.* New York: Hart, 1960.

ODELL, C. W. Marks and marking systems. In W. S. Monroe (Ed.), *Encyclopedia of educational research.* (Rev. ed.) New York: MacMillan, 1950. Pp. 711-717.

RING, K., & KELLEY, H. H. A comparison of augmentation and reduction as modes of influence. *Journal of Abnormal and Social Psychology*, 1963, *66*, 95-102.

WINER, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962.