DOCUMENT RESUME

ED 047 009 TM 000 383

AUTHOR Pemberton, W. A.
TITLE The Grade Point Average: Snark or Boojum?
INSTITUTION Delaware Univ., Newark.
PUB DATE Sep 70
NOTE 42p.

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Ability Identification, Academic Achievement, College Environment, College Students, Creative Development, Criterion Referenced Tests, Equivalency Tests, *Evaluation Methods, *Grade Point Average, *Grades (Scholastic), Multiple Choice Tests, *Predictive Ability (Testing), Predictive Validity, Sex Differences, *Student Evaluation, Success Factors, Testing
IDENTIFIERS *University Impact Study (Delaware)

ABSTRACT

This paper is a review of opinion and research concerning the objectivity and relevance of grades and grade averages as measures and as predictors of success. As measures they are ambiguous, reflecting differences in sex, basic temperament, instructors, departments, institutions, as much as levels of competence. And as a predictor of "success," grade point average has not been particularly valid for either graduate school or occupation. Criticisms of current practices--classified as ethical, rational, and pragmatic--and possible alternatives are discussed. Four suggested innovations are (1) pass-fail grading, (2) credit by examination, (3) criterion-referenced teaching and evaluation, and (4) procedures for evaluating creative extracurricular achievements. Research on the revision of evaluation procedures at the University of Delaware over a ten-year period is reviewed; and a study of the senior class of 1969 is reported in a comprehensive set of tables. The results suggest that the university has been discriminating against students who are (a) male, (b) enrolled in the sciences and traditional academic disciplines, and (c) academic nonconformists. (CK)

THE GRADE POINT AVERAGE: SNARK OR BOOJUM?

The Case for Revising
Student Evaluation Procedures

W. A. Pamberton, Ph.D.
Counseling Psychologist

A special report to the faculty which summarizes educational research and opinion on the subject of grades and grading, with particular reference to the University of Delaware.

Distributed to all faculty members

September 1970

The Student Counseling Service
University of Delaware
Newark, Delaware

For, although common Snarks do no manner of harm,

Yet I feel it my duty to say

Some are Boojums--

--Lewis Carroll, The Hunting of the Snark

THE GRADE POINT AVERAGE: SNARK OR BOOJUM?

The Case for Revision of Student Evaluation Procedures

INTRODUCTION

At the 1970 convention of college student personnel administrators in Boston, President Wofford of Bryn Mawr advised college teachers and administrators to become "students of the American college student." The concern of that conference, as it would almost have to be, was with the "growing estrangement between teachers and students" and with the "massive indifference, even antagonism, toward 'mere scholarship' on the part of student militants." (Chronicle of Higher Education, 4/13/70.)

While the initiative seems now to be with militant students and radical faculty, it seems evident that these groups cannot establish leadership in educational reform, because of increasing factionalism and internal dissent. The time is right for new initiatives which can respond to criticism without accepting the "solution" of scrapping the whole system. Some believe that if these initiatives are not taken, we may have to write an epitaph for higher education similar to that for the whole hip and psychedelic scene in the film Easy Rider: "We blew it."

Requests and demands are heard everywhere for revision of procedures in admissions, student governance, teaching, and evaluation. The concern of this report is with student evaluation, although reforms in this area would have important effects in the other areas as well. Nothing is more important to self esteem than the judgments made of one by important others. It is important that such evaluative judgments be made from data which are as relevant, comprehensive, and objective as possible.

Grade and grade averages are the only "objective" statements concerning a student's competence that appear on official transcripts. It is tacitly assumed that grades are valid for predicting success in graduate school, employment, etc., but some disturbing things are being turned up lately, which require us to question both the objectivity and the relevance of grades as measures and as predictors.

Rather than being an objective appraisal, the grade point average turns out to be quite ambiguous, reflecting differences in sex and basic temperament, and differences in instructors, departments, and institutions, quite as much as differences in level of competence. Moreover, the grade point average has not shown up as a particularly valid predictor of "success," either in occupations or in graduate school. Hoyt, after reviewing the research literature, reported that there is no convincing evidence of any significant relationship between grades and adult accomplishment. Virtually all graduate and professional schools have come to recognize that grades from different institutions and disciplines cannot be equated, and require some additional external measure for appraising candidates.

The critics are numerous and respected enough, and the evidence is convincing enough, to cast strong doubt on the canonical status of the grade point average, however respected the institution and however able its faculty.

GENERAL CRITICISM OF GRADES AND GRADING

We can pass over for now the criticisms of radical anarchists who assert that no one has the right to evaluate anyone else, at any time, in any way. Such critics, by and large, keep their fingers crossed with the reservation: "No one, that is, except me." We could agree, however, that no one has the right to tell anyone else that he is worthless.

The major criticisms of grading in colleges can be organized under these rubrics: (a) Ethical--the grading system is in part responsible for problems of morale and morality; (b) Rational--grading involves improper assumptions, and misapplications of the normal curve; (c) Pragmatic--the system does not work in achieving its implied objectives.

Ethical Issues

A survey by the Office of Institutional Studies at the University of Massachusetts on the status of pass-fail options at 22 colleges and universities, was prefaced:

> "Panic and frustration over grades are becoming so burdensome, many students and educators feel, that the cause of learning is being crushed." (Hewitt, 1969)

A report on grades and grading, prepared by the Learning Resources Center, University of Tennessee, concluded with this question:

> "To what extent can student apathy and indifference on the one hand and explosiveness on the other be the result of their being trapped within assorted, capricious and fixed systems of grade curving?" (Tennessee, 1966)

There is ethical confusion when grades are perceived by students as ends rather than means. A survey by Bowers of 5000 students in 90 institutions related class-room testing procedures and competition for grades to student dishonesty. A faculty subcommittee on student conduct at Grand Valley State College in Michigan reported that cheating is a result of the "unwholesome climate created by the emphasis upon grades." (Tennessee, 1966)

The Student Government Association of the University of Delaware conducted in 1969 a Course Evaluation survey as "a service to students and faculty," which included this question: "Did the instructor's method of evaluation provide a proper measure of your knowledge?" Of 455 courses evaluated, 282 (62%) received an average rating of "satisfactory" and 14 (3%) "very accurate" on this question; 76 (17%) were rated "poor"; and 83 (18%) were rated "Don't Know."[1]

The College Student Questionnaire, a standardized national survey, was completed in 1967 by graduating seniors at the University of Delaware. Some questions concerned attitudes toward grades and grading practices. Seniors

---

[1]SGA Course Evaluation 1969, analysis furnished by computing center.

in those departments which had the highest grade averages seemed to be most dubious about the validity of grades. Students in such "high-grading" departments, while expressing satisfaction with their own academic standing: (a) less often declared a strong interest in their major field courses; (b) more often stated that in their major departments grades were influenced by irrelevant and extraneous factors; (c) more often stated that their departments tended to reward conformity, and (d) were less satisfied with the degree of academic honesty they saw around them. (See further, page 24-27)

Students with whom I have spoken may agree that for themselves grades are not all-important, but add, "Still, it is the grade that counts--no other record appears on your transcript." When the grade becomes the end, and when grading procedures are perceived, rightly or wrongly, as irrelevant, subjective, and arbitrary, it can be predicted that students will feel justified in using devious means to achieve grades.

Florence Geis, of the Department of Psychology, has done research with the Machiavellian Scale, which measures the tendency to manipulate others for personal advantage. Dr. Geis reports that "high-Mach" students tend to make higher grades than "low-Mach" students (Geis, 1969). As an example of Machiavellianism in operation, one small class, taught by a professor known to assign grades strictly "on the curve," is reported to have pooled its resources and hired another student to sit in the course, do nothing, and take the inevitable F.

Other critics have referred to the lowering of ideals and self-respect among professionally-oriented students who are required to accomodate themselves to the "grade-making rat race." The study "Boys in White" by Becker, et al, deals with the problems of medical students who entered with high ideals and the desire to "learn everything about their profession." Most of these students had to suspend their academic ideals and adopt somewhat cynical and short-range views ("learn what you need to pass the course") in order to pass their undergraduate requirements. Only as upperclassmen were they able to reassume a more scholarly and professional perspective without hampering their achievement (Reitz, 1970).

David Riesman, with a grant from the Carnegie Corporation, investigated the problems of higher education during the period 1955-60, and reported:

> Students have often told me that it doesn't pay to be too interested in anything, because then one is tempted to spend too much time on it, at the expense of that optimal distribution of effort which will produce the best grades--and after all, they do have to get into medical school, keep their scholarship, and "please the old man."
>
> I am convinced that grades contaminate education--they are a kind of currency which, like money, gets in the way of students' discovering their intellectual interests...(Riesman, 1961).

Davis and Thistlethwaite have both studied the effects of college grades on student aspirations. Both found that with similar ability (National Merit Scholarship holders) college grades varied inversely with the quality of institution attended. Davis, for example, found that 70% of these high-ability students achieved B averages or above at low-ranking institutions,

but only 36% did so in the most selective institutions. Consequently, a smaller proportion of Merit Scholars chose to pursue graduate study who attended the more selective institutions. Their academic aspirations were lowered, because they had been evaluated by local standards. (Tennessee, 1966)

A similar finding was reported by the University of Delaware Impact Study. Two groups were contrasted: those achieving significantly higher grades than predicted, and vice versa. Among the "underachieving" seniors, 75% had planned for graduate school as freshmen, but only 13% still planned to do so as seniors. Among the "over-achieving" group, on the other hand, the percentage of those planning for graduate study had increased. The graduate plans of these students were determined by grades earned, rather than general knowledge, for on the Graduate Record Examinations the underachieving group (mainly men in engineering and physical science programs) appeared to be better prepared for graduate study. (Pemberton, C.F., March 1969)

There is no doubt that teacher-student relationships are complicated by the judging aspects of the teacher's role. The student who gets through by "out-witting the system" can hardly have respect for the system. Another type of student may avoid close personal contacts with faculty out of fear that his motives may be suspect. Thus, one student at a "Gilbert Gab" session in 1969, said: "Students hesitate to walk and talk with a professor for fear they will be accused of brown-nosing for grades."

Rational Issues

In general, grades tend to be assigned in terms of a normal distribution, somewhat independent of performance. The grades given by an individual instructor usually indicate the rank order of the student in that class, and not his performance by some broader criterion. There seems to be a conviction that "grading on the curve" is somehow more scientific.

In a survey of some 300 institutions, Benno Fricke of the University of Michigan found that the distributions of freshman grades were quite similar in highly selective institutions and those which were not selective in their admission policies. At Berkeley in 1965, for example, 30% of freshmen had grade averages below C, even though all Berkeley students came from the top 12 percent of their high school graduating classes. Hills reports that at one Georgia college the mean grade point average decreased as the institution became more selective in its admission policy. A similar paradoxical situation was observed at the University of Tennessee. (Tennessee, 1966)

The above examples can be multiplied. There are constraints within a university which almost insure that an individual instructor will not depart too far, for too long, from the local norm in assigning grades. One who habitually gives too many A's, or too many F's, usually is persuaded, one way or another, to "get with it." Grades are purely relative; an A at one institution, or in one department, is not necessarily comparable to an A in another academic context.

There has been in recent years a marked reorientation in scientific thinking about the nature of human ability, and about the most effective ways to help students learn. Samuel T. Mayo, of the National Council on Measurement in Education, says that prevailing methods of instruction and

evaluation "have promoted unsound effects on learners." The normal curve, he says, has been "overused and misused in evaluation." (Mayo, 1970)

Most people assume that mental ability and academic achievement (grades) are tied closely together. The correlations are quite high, which is not surprising since "intelligence tests" and "scholastic aptitude tests" are made that way. Trial test items which do not discriminate between educational levels are discarded in the standardization process. So, it is reasoned, post hoc, ergo propter hoc: if grades accurately reflect ability they should be normally distributed, since aptitude itself shows such a bell-shaped distribution in the general population.

The errors here are obvious. In the first place, a correlation does not imply cause-and-effect, but only co-relationship. In the second place, few, if any, college classes are random samples of the total population. In a selected sample, as represented by college students particularly at higher levels, grades would not be expected to fall into a normal curve pattern, or at least such a distribution could not be justified on the assumption of normally distributed aptitudes. Grades are based on a local population--the college, the department, or the particular class. The grade specifies a local rank order and not an absolute value; it specifies the pace of learning, but not the amount of learning in a larger world context. There is no rational basis for expecting a normal distribution of grades in a particular class unless it can be demonstrated that it is indeed a "normal" class.

A more fundamental criticism has to do with the nature of intelligence or aptitude itself. Intelligence tests and scholastic aptitude tests are constructed so as to discriminate between grade levels and age levels. As such, they necessarily are measures of precocity; the score is related to the amount of time required to achieve mastery, and not directly to any intrinsic and ultimate capacity to achieve mastery. If students are assumed to be normally distributed in that kind of aptitude, then by allowing time for achievement to vary, a greater proportion of students can be expected to achieve mastery. This, essentially, is Mayo's argument, when he endorses "criterion-referenced" as opposed to "norm-referenced" testing and evaluation. In criterion-referenced learning, the student is evaluated on how many steps he can take, not how fast he steps, nor what percent of the group is in front or behind.

Such a concept has far-reaching social implications. Rather than thinking of aptitude as a kind of ceiling, which is imposed genetically, we regard aptitude as being developed over time. The fact is, as Stoddard has been saying for 30 years, that people have to learn to be intelligent, or learn not to be, for that matter, and some learn faster than others and in different ways (Stoddard, 1943). More people can achieve a high level of intellectual functioning than has been thought. The traditional ways of teaching and evaluation, in school and in college, have succeeded in teaching some students that they cannot become intelligent. In this the system has been educationally and socially counterproductive.

The experiments of Rosenthal and Jacobson showed that intelligence test scores could be raised when teachers were told "that is what the tests say", even though the test data were spurious. When pupils were enabled to think of themselves as bright, because teachers acted toward them as if they were,

they achieved higher intelligence test scores. The "self-fulfilling prophesy" is a real psychological phenomenon which complicates greatly our ethical and scientific responsibility in evaluation. (Rosenthal, 1968)

The experiments by Krech, Rosenzweig, et al, at California (Berkeley) during the past 15 years have proved that measurable and significant changes in brain structure and brain chemistry can be brought about (at least in laboratory animals) by social and intellectual stimulation. It is becoming increasingly untenable as an educational hypothesis that intelligence and other aptitudes exist as fixed and unchanging entities. (Rosenzweig, 1966) Teaching objectives rationally should be based on the premise that all students can learn, although at different rates. New methods, Mayo says, may be required to bring this about.

> Whereas reward and punishment (read this "grades") were once paramount in theories of learning, ideas of organization and structure now dominate major innovations such as programmed instruction and computer-assisted instruction. (Mayo, 1970)

The Pragmatic Issue

The crux of the matter, of course, is whether grades work in achieving their explicit and implicit functions. There must be a reason if, in spite of strong negative criticism, the system continues to survive. It is not easy to find published articles which actively support the grading system, even though we all know that there is strong and responsible opinion on the pro side of the argument.

Paul Goodman says that the retaining of grading in colleges is "an interesting case of bureaucratic inertia and subservience to the social climate." Although many teachers agree that the grading function hurts teaching and learning, still they see it as inevitable because of extramural pressures from employers, graduate schools, and parents, as well as from students themselves. (Goodman, 1964)

Grade for employers. The fact of graduation in a particular program usually provides sufficient basis for employer acceptance of applicants. If more evidence of specific skills and attributes is required, that kind of evaluation is more appropriately done by the employer than by the university. The areas of civil service, accounting, medicine, etc., all provide their own tests for licensing. Moreover, the grade point average does not offer dependable information to employers. Hoyt, after reviewing some 50 research studies which related college grades to adult achievement in various fields, concluded:

> The present evidence strongly suggests that college grades bear little or no relationship to any measures of adult accomplishment. (Hoyt, 1965)

Now, it would be neither fair nor accurate to infer from Hoyt's study that grades mean nothing. By the time of graduation from college, students are highly selected for ability, as a rule, and the small degrees of difference between "A" and "B" students do not represent crucial differences in human ability and human worth. Moreover, such positive relationships

between grades and adult achievement as might exist are obscured by throwing together grade averages from different institutions or from different departments within the same institution. Similar grade averages represent different kinds and levels of achievement in differing educational contexts, and are not really comparable. To say that one is a "B student" without further specification is much like saying that a student "scored at the 75th percentile on a test" without specifying the test or the norm group used for comparison.

The criterion or standard of "success" is another sticky problem when evaluating the significance of grades. For example, we do not need a Texas millionaire to tell us that higher education is not a necessary prerequisite for becoming wealthy. However, the world has the right to expect from us that college grades should be correlated with something besides the ability to make grades. We should provide an additional objective judgment, something external to the grade-making situation, if we wish to have our evaluations of students taken seriously.

Grades for graduate schools. Colleges and universities recognize the inadequacies of high school grades for determining admission and placement, and most of them now require supplemental evidence from nationally-normed tests. It seems only rational that the colleges should recognize the limitations of their own grades for predicting performance at the next step. Even if the undergraduate colleges do not admit it, those who accept our graduates do recognize the erratic nature of grades and usually insist on some external standard of evaluation which will help to equate institutions and disciplines.

Lannholm's report on studies conducted by himself and others casts doubt on the validity of undergraduate grade point average for predicting success in graduate school The single best predictor of pass-fail success in graduate school, he says, is the Advanced Test of the Graduate Record program; and a combination of the GRE Aptitude Test and Advanced Test consistently yields more valid prediction of graduate school success (at the Ph.D. level) than undergraduate grades. When such an external measure is not available to the graduate schools, most of them resort to some device for "weighting" the grades from particular institutions and departments. (Lannholm, 1968)

Grades as motivators. It is said that students will not work unless graded; that grades are needed as an extrinsic spur. There is evidence that some students learn less under a pass-fail system than when letter grades are used. While there are such students, dependent on external structure and outside pressure for their motivation, there are other students who appear to accomplish more when permitted to proceed at their own pace. For some students, learning is inhibited by lock-step instruction and having to work for grades. The argument comes down to this: grades and threats of flunking a student are the only whip that teachers have left for keeping lazy and unruly students in line. This is a losing battle. If one cannot show students that he is an authority on something, he will not long be able to exercise authority over them.

In his Theory of Instruction, Bruner is especially critical of the theory that the will to learn can be imposed on students from without, and

of the pedagogical model which depends on extrinsic rewards and punishments for its motivating force.

> The will to learn is an intrinsic motive, one that finds both its source and its reward in its own exercise. The will to learn becomes a "problem" only under specialized circumstances like those of a school, where a curriculum is set, students confined, and a path fixed....What the school imposes often fails to enlist the natural energies that sustain spontaneous learning--curiosity, a desire for competence, aspiration to emulate a model...(Bruner, 1966, p. 127)

The will to learn, Bruner says, and the most satisfying state of affairs from the viewpoint of teacher and student, are not reliably to be achieved by kind or harsh words from the teacher, by grades and gold stars, or by the "absurdly abstract assurance to the student that his lifetime earnings will be better by 80 percent if he graduates."

Grades and Creativity

Paul Goodman. Critics such as Goodman say that competitive grading, with the counting of credits and quality points, and the assembly-line speed-up are part of a "cash accounting and logistic mentality" that is not conducive to the development of creative students. (Goodman, 1964)

Jerome Bruner. In his Process of Education, Bruner warns against grading. He says that, particularly in the science fields, students should be taught, not facts, but ideas and methods. It might on occasion be more important that a student spend a whole term checking out why his experiment did not work--a project that is essentially not gradable.

W. J. Bender. The former dean of admissions at Harvard is quoted:[1]

> A study of the truly creative and original Harvard graduates would, I believe, reveal only the loosest correlations between school and college records and subsequent attrition, and a low proportion of summa cum laude graduates among the gifted.

Sidney Harris. This syndicated columnist wrote:[2]

> Every study made of achievers in a genuinely creative sense--people who were truly innovative--has shown that as children these people were anything but docile and conformist...Many genuine achievers such as Edison and St. Thomas Aquinas (he could have added Churchill, Einstein, and others) received distressingly poor grades in school.

L. L. Thurstone. Thurstone, one of the great psychologists, used to say that when selecting a graduate assistant to work with him in his Psychometric

---

[1]Intercollegiate Press Bulletins, January 1, 1962.

[2]Wilmington News-Journal, March (?), 1970.

Laboratory, he tended to prefer one "who had enough imagination to have failed a course in college."[1]

David Riesman. After trying "in vain" to persuade students that they could think less about grades and more about education, Riesman says that one of his graduate students at Chicago finally did a thesis which documented his arguments. The student asked departments which graduates they had recommended for jobs, advanced training, fellowships, etc., and then interviewed the students and looked at their grades. It seemed that those students fared best who were "not too obedient" and did not achieve straight-A records. The most highly recommended students were "a bit rebellious and off-beat, although not goof-offs." Students, to be sure, had to do something to earn these commendations, but they were usually better off to have done something unusually well than to have opportunistically allocated their time so as to achieve highest grades in all courses. (Riesman, 1961)

Phi Beta Kappa. A committee was appointed in 1967 by the national president of Phi Beta Kappa to consider the implications of ungraded courses and the evaluation of such courses when appraising candidates. A poll of chapters revealed ambivalent opinions as to the educational advantage of the pass-fail option as a solution, but general agreement as to the problem of insuring that the Phi Beta Kappa key is not a "badge for grinds." The committee reported:

> The investigations of the Pass-Fail Study Committee revealed that many chapters place far too much emphasis on the grade point average in the selection of members. The Committee recommends that in the selection of members due attention be paid to factors other than the grade point average, such as evidence of genuine intellectual interest and distinguished scholarship. (Phi Beta Kappa, 1965)

Grading Interferes with the Proper Functions of Testing

The point is made by Bruner, Goodman and Mayo that when tests are used for grading, an important educational function is destroyed--that of diagnosis. Bruner says that the educational experiment, in the main, is being conducted in the dark, without usable feedback. The substitute for feedback is evaluation after the job is completed: "after the working party has been scattered, the evaluators enter." It would seem more sensible, Bruner says, to provide evaluation during the operation, so that errors can be corrected. (Bruner, 1966, p. 163-4)

Mayo recommends that teachers use class quizzes for evaluation of progress, not for grading. A final qualifying level of competence could be determined by comprehensive examinations, which could be repeated in alternate forms.

One of the most important aspects of programmed instruction is immediate feedback, which provides reinforcement at the optimal time for learning. When tests are used to determine grades, however, test items often are used over

[1]Personal Communication.

and over, the examinations must be kept "secure," and the student is deprived of the opportunity to learn from the test.

There are proper functions of examinations, Goodman says, which are inhibited by grading.

> We properly examine a candidate to see if he is acceptable into an enterprise or community. Once he is in, why distinguish one from another in the class, like first and second class citizens?...A second proper kind of examination is to see if the youth has now grown up to be a peer. In medieval times, he proved his entry into the guild by a masterpiece--academically, a lecture and disputation. The point of this, however, is to do something that wins respect--not to pass somebody else's questions, which would maintain him precisely in his condition of inferiority and immaturity. (Goodman, 1964)

The Berkeley and Carnegie Reports on Grading

A report of the Select Committee on Education at Berkeley devoted a chapter to the problem of grading. Prof. Stewart Miller reviewed the literature, conducted interviews with students and faculty, and corresponded with other colleges and universities. He reported that only a bare majority of students believed in the efficiency of the system, with objections to grading not being confined to those graded low. Faculty opinion, pro and con, showed the same range. Those defending the system spoke of the "unwelcome but salutary comparison by which each student is forced to learn something about his standing among peers, and also to criticize himself." In a letter to the committee, Prof. Thomas Nagel argued:

> It would be deplorable if the rather harsh, critical environment appropriate to an educational institution gave way to a congenial, unevaluative one, in which scholars went about their business and students were simply welcome to pick up what they liked, as spectators on the intellectual scene.

Another letter to the committee from Prof. Henry May, was more critical of the system.

> I have very little confidence in the grades turned in as a result of examinations read by 20 different teaching assistants, many of them grading for the first time, in a class of 1000. Such large numbers put a premium on the easiest and most efficient methods of grading, rather than the most serious ones.

The Berkeley Committee in its final report did not recommend wholesale revision or abolition of the grading system, but recognized some of its weaknesses.

> At the same time we cannot express satisfaction with the actual conditions under which we grade on this campus; and there are grounds for challenging them on two issues, with respect to both how we grade and the use to which grades are put. By the latter we mean our ubiquitous calculation of grade point average as a criterion for academic privileges, including honors standing and advancement to graduate study.

There is no doubt that an obsession with grade points drives many of our students to choose courses for perverse reasons.

In its summary the Berkeley Committee recommended: (a) that teachers should grade less often in order to grade better; (b) that the principle of counting all units equally in calculation of the grade point average should be seriously questioned; (c) that first-term grades should not be counted in the grade point average, although counted toward graduation and (d) that considerations other than grade point average should be taken into account in allocating honors and special academic privileges. (Berkeley, 1966)

The Carnegie Commission on Higher Education has recently surveyed more than 60,000 faculty members on various educational issues. One statement read: "Undergraduate education would be improved if grades were abolished." The faculty response was: Agree 31%, Disagree 66%, No Response 3%.[1]

Clearly, any proposal to abolish grades would meet with overwhelming opposition from college faculties. Just as clearly, some reforms in present grading practices are favored, and required. Richards (1970) after reviewing research in the field was obliged to conclude that "all is not well with current methods of assessing student accomplishment."

[1]Reported in University of Delaware News, Spring 1970, p. 35.

STUDIES CONDUCTED AT THE UNIVERSITY OF DELAWARE

The case for revision of student evaluation procedures at the University of Delaware has been built up over the past decade, starting with the inception in 1960 of the Commission to Study the Impact of the University on its Undergraduates, now termed The Impact Study.

Some say that it takes a new idea 10 years to catch on in education, and another 10 years to be implemented, due in part to inertia, and in part to the opposition of those with vested interests in the status quo who wield more clout. A more important reason, probably, is that new ideas, even if sound, are seldom well enough understood, explained, and communicated. The points need to be made over and over again, in different ways.

The University Impact Study

A monograph was published in 1963 under the aegis of the Impact Study entitled Ability, Values and College Achievement, which evaluated from several aspects the class graduated in 1960 (Pemberton, W.A., 1963). This study showed that grade point average identified "educationally conforming" students, but not necessarily the most creative, nor in all cases the best educated. A factor analysis showed two distinct kinds of achievement: (a) General College Ability, defined by external tests; and (b) Academic Achievement, identified by high school and college grades. Students identified with the grade-making factor more often were (a) women, (b) students with practical, vocational goals, and (c) students who scored high on measures of social and academic conformity. Conversely, students showing up on the factor defined by tests were more likely to be (a) men, (b) students enrolled in liberal arts programs, and (c) those less conforming in temperament.

The study recommended that student competence should be defined by a three-fold criterion: grade point average, external examinations, and some evidence of creative or independent production. Partly as a result of this study, the Commission recommended that the Graduate Record Examinations (Area and Advanced Tests) henceforth be required of all graduating seniors. The recommendation was endorsed by Paul Dressel of Michigan, consultant to the Commission; and was given impetus by the statements of Cyrus Day of the English Department, then chairman of the committee to select local candidates for the Woodrow Wilson Foundation fellowships. Professor Day in a letter to the faculty described himself as "becoming embarrassed," as some departments continued to nominate their "best" students (as determined by grades) only to have them turned down by external examiners as parochially educated. He recommended that the faculty look beyond grade point average when nominating fellowship candidates.

A report of the University Impact Study in 1966 evaluated the grade point average as a means of identifying superior students. This report showed that grade point average underestimates the general cultural knowledge of students in certain curricula, overestimates the general competence of students in others, and generally favors women while discriminating against men (C. F. Pemberton, 1966).

Another report in 1969 contrasted two groups of graduating seniors: (a) those achieving significantly higher grades than predicted, and (b) those with

high-ability (tests) but relatively low grades. Temperamental and motivational characteristics of these two groups were compared, based on the College Student Questionnaire taken by these students as freshmen and again as seniors. The higher-test students were found to be more independent, self-directed, and culturally sophisticated. The higher-grades students, on the other hand, were described as relatively more dependent, conforming, and narrow in their interests. (C. F. Pemberton, March 1969)

The above report, it should be clear, dealt with two extremes. The most competent students will usually do well on both measures, tests and grades. It was implied, however, both by this and the earlier studies, that if either method of evaluation were to be devalued, we might be wiser to deemphasize grades than the external examinations.

Three reports of the Impact Study analyzed the class graduated in 1969 in terms of ability, values, and attitudes. One report released in 1970 dealt with grades and two external examinations: (a) the Graduate Record Area Tests taken as seniors, and (b) the College Level Examinations, General Battery, taken in the sophomore year. Although tests and grades were positively correlated within each department, when University-wide comparisons were made the departments were rank-ordered differently by test averages and by GPA averages. That is, while grades reflected some general consensus as to a student's competence, grades also were determined in part by the particular department in which the student was enrolled. It was concluded that grades provide a more limited appraisal of the student, since "each instructor uses his class as its own norm and does not compare the students in that class with the total University population." Students in Arts and Science and Engineering were shown to be under-evaluated by grades; student in Business and Economics were evaluated equally well by external tests and grades; and in Agriculture, Education, Home Economics and Nursing, grades were high, relative to test scores. (C. F. Pemberton, March 1970)

Two other reports by the Impact Study, not widely enough circulated and appreciated, establish that ranking academic divisions by grades corresponds closely to rankings for general conformity and acceptance of the educational status quo. There also is a tendency for students in those departments highest in grade average to be more dubious about the intellectual integrity of the grading system. (These reports are discussed further, p. 24-7.)

These various reports of the Impact Study seem to show that the present reward system of the University of Delaware operates in such a way as to penalize students for being (a) male, (b) enrolled in Arts and Science or Engineering programs, and (c) nonconforming in temperament. In the process, cynicism is being encouraged. As one student put it, "Cooperate and graduate--that's the way the game is played here."

Other Local Studies

A particular research interest of Marvin Zuckerman, Department of Psychology, is the personality variable that he calls "sensation ~~stimulus~~-seeking." Persons scoring high in SS actively seek out new experiences, and are more prone to rebellious behavior and skeptical attitudes. Men typically score higher than women on SS. In one study, Zuckerman's test was found to be positively correlated with scholastic aptitude test scores and with measures of general intel-

ligence, but not with grade point average. Zuckerman believes that the potential advantage of ~~stimulus~~ sensation-seeking temperament for self-directed learning is counterbalanced in the grade-making situation by the disadvantages accruing from nonconformity.[1]

An ad hoc subcommittee of the Committee on Academic Status of Undergraduates (CASU) was appointed in 1968 to study the matter of course credit by examination. The subcommittee, comprising Ronald Wenger (Chairman), Reuben Austin, Bessie Collins, E. W. Comings, Robert Mayer, and W. A. Pemberton, gave its report in October 1969. This report recommended (a) that CASU adopt the principle of awarding course credit by examinations prepared outside or within the University; and (b) that a student's scores on the Graduate Record Examinations (Undergraduate Program) be required and placed on his permanent record before graduation.

The subcommittee in its meetings considered some of the broader questions related to grades and grading, and reached the following conclusions which were presented in an interim report dated May 21, 1969:

1. Women tend to make higher grades than men, even though scores on standardized achievement examinations may be lower.

2. Average grades differ from one course of study to another, so that graduation, honors, and admission to graduate school are determined in substantial degree by the course of study pursued rather than over-all level of educational development.

3. Conforming students tend to make higher grades; while independent, creative students of the same or higher ability tend to make lower grades. The latter students are discriminated against in terms of graduation, index requirements, honors, and admission to graduate school.

The evidence on which the above report of the faculty committee was based, together with more recent supporting data, is presented in some detail in a following section, page 18f.

---

[1]Personal communication, forwarding excerpt from an unpublished study by Kish, G.B. and Dannenworth, G.V. on ~~Stimulus~~ Sensation Seeking relationship to capacity and personality.

EXTERNAL EXAMINATIONS

Frequent reference has been made in this paper to "external examinations," particularly the Graduate Record Examinations. In view of reservations held by some faculty members concerning these examinations and concerning "multiple-choice tests" in general, some further explanation and defense seems required at this point.

An expert in teaching is not necessarily an expert in testing or test construction. A teacher seldom has the time, even if he has the skill, to "test the test" that he has constructed. Not many universities have staff or facilities to operate a competent Office of Examinations such as those at the University of Chicago and the University of Michigan. Under these circumstances, the Graduate Record Examinations (Undergraduate Program, for appraising senior-level competence) and the College Level Examination Program (for freshmen and sophomores) can be of help, locally.

The Graduate Record Examinations--Undergraduate Program, presently required for graduating seniors, includes the Area Tests of general education (Humanities, Social Science, and Natural Science) and Field Tests in 28 specific course areas. The College Level Examination Program includes a five-test general battery, together with Subject Tests presently available in some 30 areas, a number which eventually will be increased to 100. (The CLEP general tests now are required for all students transferring to the University of Delaware.) For both programs, sample questions are provided for students, and item analysis services are available to teachers for diagnosis of special strengths and weaknesses in student achievement.

The examinations are constructed by the Educational Testing Service of Princeton, New Jersey, a non-profit educational enterprise with an impressive reputation for ethics and competence in educational research and test development. While objective in form, the tests are devised to emphasize understanding of broad principles, and the ability to apply these principles to new problems. (Optional essay forms are available for the CLEP Subject Tests) Publications of ETS have dealt with the "recurring myth that the multiple-choice question is a superficial exercise that requires little thought, less insight, and no understanding." Like other myths, this one is based on shadowy memories of one's own experiences (perhaps True-False examinations constructed by elementary teachers) and bears little relation to present reality.

Jerome Bruner, in The Process of Education, says that the objective-essay issue in testing is irrelevant:

> Whether an examination is of the 'objective' type involving multiple choices or of the essay type, it can be devised so as to emphasize an understanding of the broad principles of a subject. (Bruner, 1960)

Either type of examination, essay or objective, can measure trivia if that is what one chooses to test for. For those who are dubious about objective-type examinations, a review of the examinations would be informative. The pamphlet Multiple-Choice Questions: A Close Look (Educational Testing Service, 1963) provides sample questions from a variety of tests published

by ETS. A close look at this representative sample of multiple-choice questions should lead to a better understanding of their potentialities, and help dispel the myth that objective tests require no thought, insight, or understanding.

The "ceilings" of the tests are high enough that even the brightest students are challenged; and the content range is wide enough that highly specialized students can find some areas in which to demonstrate their special competence.

Virtually every department of the University which during the past 10 years has presented its "self-evaluation" in the Impact Study, and more recently its long-range plans before the Community Design Commission, has emphasized its "general culture" as well as "practical skill" objectives. And, no department has been willing to base its case on a purely internal evaluation, saying: "Look how many A's (or how few A's) we give in our department." The departments recognize that if their evaluations of themselves and of their students are to be recognized by their academic peers, performance must be validated against some external standard. Typically, the departments have found it more convincing to be able to point to student performance on the Graduate Record Examinations; performance on professional "boards" set by law, medicine, accounting, etc.; or to the number of students admitted to and successful in graduate school, which itself usually is contingent upon the external examinations.

These examinations are not equally relevant for all curricula and all courses. There are some areas of competence for which a paper-and-pencil test is not an appropriate evaluation medium. Competency in a course might sometimes be determined better by practical demonstration, laboratory exercise, practicum experience, or personal interview than by conventional tests. For all students receiving baccalaureate degrees, however, it seems entirely appropriate that there should be some standardized evaluation of general cultural and educational development.

Boozer examined the case pro and con for external examinations and concluded:

> While there are valid arguments against external examinations as predictors of competence, they are not so persuasive as are those in favor of the intelligent use of such examinations...They enable the colleges to view their graduates...in relation to persons of comparable educational exposure throughout the country. (Boozer, 1965)

The strongest arguments against the use of external examinations, Boozer concluded, were (a) that some skills and proficiencies cannot adequately be tested by paper and pencil, and (b) that some persons might use such tests to advance pet schemes of their own, allowing the tests to assume such prominence in education that they become "the tail that wags the dog." The strongest argument for such examinations, Boozer decided, is that "the use of proficiency examinations...will lead to more flexible programs and to the encouragement of student initiative and self-directed work."

Paul Dressel critically examined the role of external testing programs, and after reviewing complaints and fears regarding such programs, concluded

that the objections were insufficient. The major concern, he found, was that external examinations constitute a threat to local autonomy:

> ...college faculty and administrators pride themselves on the individuality of their programs. Faculties...resent the second guessing of judgments already rendered on the accomplishments of their students. (Dressel, 1964)

In most institutions, Dressel concluded, "neither the pride nor the resentment is justifiable." There is in general, he said, greater uniformity in course offerings than the claims to uniqueness would suggest, and most programs would profit, not suffer, from being required to submit their records for external auditing.

It should be feasible at this university, at least in some areas, to use teacher-made tests for teaching, and external examinations for final evaluation. J.M. Richards, in a review of research for the Eric Clearinghouse on Higher Education, considers the College Level Examination Program a promising method for enabling individuals who have acquired knowledge in non-traditional ways to demonstrate their academic achievement. While the CLEP general examinations overlap to a considerable degree the characteristics measured by general scholastic aptitude tests, the CLEP subject tests seem well designed to serve as comprehensive examinations, and as ways to permit students to obtain credit by examination.

Richards' reservations concerning the CLEP general examinations have to do with their high correlations with traditional scholastic aptitude measures, and not with the content and structure of the tests themselves. This objection may reduce to one of epistemology--there simply is not that much difference in "aptitude" and "achievement" measures. Whether we want to call the measure an aptitude or achievement test depends largely on whether we want to predict future achievement or measure past achievement. We can do either with the same test. The common factors in the SAT, GRE, and CLEP are probably Reasoning and Verbal Ability, which are required for most paper and pencil tests.

A recent report by the University Impact Study showed clearly that the sub-tests of CLEP and the GRE Area Tests provide measures of specific achievement as well as relative competence in all areas. This is better demonstrated by graph (Figure 2, page 20 of referenced report) than by correlations. There were significant differences among curriculum areas on average scores (i.e., high intercorrelations, since bright and well-read students do better on all tests) but there were also highly significant differences (one to three stanine intervals) between sub-tests for most curricula, with highest mean scores being associated with the appropriate major field. While the CLEP and GRE composite scores were highly correlated with each other, the relative gains from the sophomore to the senior year were greatest in the area of academic concentration. (Pemberton, C.F., March 1970)

EVALUATION OF THE SENIOR CLASS OF 1969

Sex Differences in Grades

It has long been observed that girls make better grades in school than boys. A report by the Delaware State Board of Public Instruction, for example, showed that two-thirds of Delaware high school students graduated in the top quarter of their classes in 1967 were girls, while two-thirds of those in the low quarter were boys. That this same pattern continues to exist in college is not always appreciated.

A Group Index Summary is prepared by the Office of Admissions and Records and distributed to administrative officers at the end of each semester. This summary shows the distribution of grade point averages by college, by sex, and by living groups. The summary of averages for sex and college at the end of the first semester in 1969 is shown in Table 1, left hand side. On the right is shown a similar summary by sex and college for class of 1969, which includes GPA for the junior and senior years only, a period during which, presumably, most grades were earned in major field courses.

Table 1. Grade Point Average by Sex and College

Cum. GPA, Sem. 1, 1968-69				Junior-Senior GPA, Class of 1969			
College	N	Women	Men	College	N	Women	Men
Agriculture	29	2.92		Agriculture	5	3.21	
Engineering	12	2.78		Education	115	3.02	
Arts/Science	1306	2.71		Home Economics	74	2.97	
Home Economics	396	2.69		Nursing	36	2.86	
Education	1012	2.58		Arts/Science	284	2.84	
Nursing	243	2.55		Engineering	1	2.76	
ALL-STUDENT AVERAGE		2.50		ALL-STUDENT AVERAGE		2.70	
Arts/Science	1594		2.42	Agriculture	61		2.70
Engineering	817		2.38	Arts/Science	262		2.66
Bus./Econ.	89	2.35		Education	20		2.61
Agriculture	421		2.35	Bus./Econ.	18	2.49	
Education	279		2.26	Bus./Econ.	107		2.44
Bus./Econ.	789		2.26	Engineering	122		2.35
Nursing	3		0.93				

When grade point averages are tabulated in this way, it can be seen that in every college enrolling both men and women, the average grades for women are higher, both in terms of cumulative grade point average and in terms of grades earned during the junior and senior years. There is hardly any overlap; except for the College of Business and Economics, the average for women

is higher in each college than the All-Student average. Separation by grade point average according to sex makes it appear that we have virtually separate colleges for men and women.

The situation is by no means unique to Delaware, and seems to be a universal phenomenon in American colleges. Yale admitted its first undergraduate women in 1970, and sure enough, the first-semester reports showed women excelling in the top grade categories, Honors and High Pass; while there was a higher percentage of men in the lower-grade categories of Pass and Fail.

Graduation

The five-year record of students who entered the University of Delaware in 1965 is shown in Table 2. There were 1489 first-admission freshmen who entered degree programs in September 1965, for whom complete freshman test records are available. The entering class was made up of 53% men and 47% women. Of this group, 39% of the men and 53% of the women were graduated four years later (46% of the entering class.) Another 16% were graduated in 1970, and 4% were still continuing in reclassified status. In the latter two groups, more than three-fourths were men. A higher percentage of men than women it appears, will eventually receive their degrees, but women are more likely than men to complete degree requirements in four years. Since women typically receive higher grades than men, this could be expected.

Table 2. Status Five Years Later of Students Entering in 1965 as First-Admission, Full-Time Degree Candidates

Entered 1965			Grad. 1968-69		Grad. 1970		Reclassified		Total Continuing	
	No.	%	No.	%	No.	%	No.	%	No.	%
Men	797	53	308	39	178	22	50	6	536	67
Women	692	47	371	53	58	9	16	2	445	64
Total	1489		679	46%	236	16%	66	4%	981	66%

Graduation with honors

Another way to study sex difference in achievement is in terms of students graduated with honors. In June 1969, 1252 seniors were graduated. Of this group, 697 (about 55%) had entered the University of Delaware in 1965; the other 45% had entered earlier, or had entered by transfer. In the graduating class, 128 (10% of the total) were graduated with Honors, High Honors, or Distinction, recognitions for which the basic requirement is a cumulative grade point average of 3.25 or above. (Students graduated with Distinction are not held strictly to this grade requirement, but must present and defend an honors thesis. Students receiving High Honors have been interviewed by a committee of external examiners.)

Table 3. Honors Graduates, Class of 1969

	Total Graduates		Honors Graduates		
	No.	% of Total	No.	% of Honors	% of Total
Men	697	56%	55	43%	7.9%
Women	555	44%	73	57%	13.2%
Total	1252	100%	128	100%	10.2%

While the graduating group was made up of 56% men and 44% women, in the honors group these percentages were reversed: 57% women and 43% men. This relationship is shown in Table 3.

Any way we look at it--cumulative grade point average, grade point average for upperclass courses only, percentage graduated in four years, and percentage graduated with honors--women show up as better students than men, when grade point average is the primary basis for evaluation. Well, maybe they are.

Sex and Curriculum Differences on Tests

The next question is whether the clear superiority of women in terms of grade-making extends to performance on external tests of competence. Three examinations can be used to make comparisons for the class of 1969, with respect to sex and curriculum differences.

1. SAT. The Scholastic Aptitude Test of the College Entrance Examination Board was submitted with application for admission in 1965.

2. CLEP. The College Level Examination Program, General Battery, comprises five tests: Humanities, English Composition, Social Science, Natural Science, and Mathematics. (Taken by sophomores in April 1967).

3. GRE. The Area Tests of the Graduate Record Examinations were taken by these same students as seniors in April 1969. The Area Tests comprise Humanities, Social Science, and Natural Science.

The above tests and test batteries were each reduced to a single composite or average score. The three composite scores were found to be highly intercorrelated (all above .70 with average correlation .75). Since each test has the same mean (500), same standard deviation (100), involve essentially the same subjects, and are highly intercorrelated, a single composite score or average of the three examinations, can be used to represent "test competence."

Table 4 shows composite test scores and upperclass grade point averages for 16 academic divisions (13 for men, 14 for women). Grade point averages, computed only for the last four semesters, represent achievement for the most part in major field courses. The group includes all students entering in 1965 who continued through the junior year and who completed at least two of the three test programs: SAT, CLEP, and GRE. Of this group, most were graduated either in 1969 or 1970. A few withdrew in the interim, or were still continuing in the reclassified status during 1970, but the group essentially comprises college graduates.

Table 4. Academic Ability as Measured by Tests and Grades

Academic Division	Test Average			Upperclass GPA			N		
	M	F	Tot.	M	F	Tot.	M	F	Tot.
Phy Sci	590	689	608	2.65	2.70	2.66	23	5	28
Math/ComS	600	585	591	2.72	2.96	2.83	27	22	49
Chem Engr	588	-	588	2.59	-	2.59	35	-	35
Bio Sci	597	568	586	2.79	2.87	2.82	51	31	82
Soc Sci	574	566	571	2.66	2.75	2.69	56	29	85
CE,EE,ME	559	(610)*	559	2.51	(2.80)*	2.51	63	1	64
Humanities	574	552	559	2.66	2.80	2.76	29	69	98
Behav Sci	560	557	559	2.52	2.77	2.64	29	26	55
Second Ed	533	544	540	2.49	2.83	2.71	15	27	42
Home Econ	--	533	533	--	2.95	2.95	--	65	65
Bus & Econ	532	525	531	2.42	2.61	2.44	111	11	122
Nursing	--	530	530	--	2.92	2.92	--	35	35
Fine Arts	546	510	524	2.65	2.66	2.66	15	24	39
Agricul	516	576	521	2.66	3.16	2.71	54	5	59
Elem Ed	(477)*	506	505	(3.05)*	2.89	2.89	2	98	100
Phys Ed	475	474	474	2.60	3.30	2.96	12	13	25
Total	557	538	548	2.57	2.86	2.71	522	461	983

*The "averages" in parentheses are based on 1 or 2 students and cannot be regarded as representative. The averages for women in Physical Science and Agriculture (5 in each group) must also be interpreted with caution.

Two things stand out from Table 4. First, there are large differences among curriculum groups, both in terms of grade point average and test average. Second, while women in each academic division excel in upperclass grade point average, men excel on the external examinations in virtually every division (the exceptions are for groups with quite small numbers). Women show up as better students when evaluated by grades; men show up better when evaluated by external tests of competence.

A further comparison in Table 5 shows the irrational effects of comparing students across the board for grade point average (as the University

does now) with curriculum and sex differences ignored. The test and grade averages shown above in Table 4 have been changed in Table 5 to ordinal ranks, and rank difference correlations have been computed. (The rank difference correlation procedure is statistically legitimate, when certain requirements of normal distribution are met, and provides a reasonable approximation of the more familiar product moment correlation coefficient.)

Table 5. Grade and Test Averages Expressed in Ordinal Ranks, and Resulting Rank Difference Correlation Coefficients

Academic Division	Rank Orders for Academic Divisions: Men Tests	Men GPA	Women Tests	Women GPA	Total Class Tests	Total Class GPA
Physical Sci.	3	6.5	1	12	1	11.5
Math, Comp S.	1	2	2	3	2	5
Chem. Engrg.	4	9	--	--	3	14
Biol. Sci.	2	1	4	7	4	6
Social Sci.	5.5	4	5	11	5	10
CE,EE, ME	8	11	--	--	7	15
Humanities	5.5	4	7	9	7	7
Behavioral Sci.	7	10	6	10	7	13
Secondary Ed.	10	12	8	8	9	8.5
Home Econ.	--	--	9	4	10	2
Bus. & Econ.	11	13	11	14	11	16
Nursing	--	--	10	5	12	3
Fine Arts	9	6.5	12	13	13	11.5
Agriculture	12	4	3	2	14	8.5
Elem. Ed.	--	--	13	6	15	4
Physical Ed.	13	8	14	1	16	1
Rank Difference Correlation	.55		-.02		-.35	

In general, we can expect tests and grades to be positively correlated within each academic division, higher test scores being associated with higher grades. As has been shown above, however, grades are local measures, specifying a student's rank within a specific group. Grades also vary with sex and curriculum. These differences are maximized in Table 5, where GPA rank is based on upper-class grades only, a period in which, presumably, most grades are earned in major field courses. If we ignore these sources of bias when making evaluation decisions, and act as if grades mean the same thing for all students in all curricula, some irrational decisions will be made.

Table 5 shows that men tend to be ranked similarly by upper-class grades and tests (Rho=.55). The exceptions are for chemical engineering (lower grade rank), and agriculture and physical education (lower test rank). (Without agriculture, for example, the rank order correlation for men would be

increased from .55 to .67.) Among women, however, there is no relationship between the rankings of curricula by grades and by tests (Rho= -.02). Women in physical science rank first in average test scores, but 12 in (upper-class) grade rank. For women's physical education there is a complete reversal: rank 1 for highest grades and rank 14 for lowest tests.

Both sources of systematic bias (sex and curriculum) are compounded when grade comparisons are made for total students in all curricula, as shown in Table 5. When this is done, we have the paradoxical situation that ranks are negatively correlated; that is, there is some tendency for higher grades (in the last two years) to be associated with lower test scores (Rho= -.35). It follows that when graduation, honors, and other academic rewards are conferred purely on the basis of grades, some students, almost certainly, will receive the academic kudos who are less able than some who are left out.

A report by the University Impact Study on this same class analyzed the relationship between grades and two external measures of achievement: GRE Area Tests and the College Level Examinations, general battery. (Pemberton, C.F., March 1970) In that study, the term "grades" referred to four-year cumulative index, not upper-class grades only as in the present study; and only 11 curriculum groups were used for comparison. Rankings are shown in Table 6, with the 16 academic divisions (shown in Table 5) reduced to 11 to make possible a comparison with the Impact Study data.

Table 6. Comparison of Test and Grade Rank Orders in Two Studies

Academic Division	Test Rank Order		Grade Rank Order	
	Present Study	Impact Study	Present Study	Impact Study
	(4-test aver.)	(2-test aver.)	Upper-Class GPA	Cumulative GPA
Phys Science	1	1	6	3
Biol Science	2	2	5	1
Engineering	3	4.5	10	6
Soc Science	4	3	9	10
Humanities	5	4.5	7	5
Home Economics	6	6	2	4
Bus & Econ.	7	8	11	11
Nursing	8	7	3	2
Agriculture	9	9	8	7.5
Elementary Ed.	10	10	4	7.5
Physical Ed.	11	11	1	9
Rho	.97		.42	

The two test rank orders are almost identical (Rho= .97). The effect of basing grade rank on an average of four examinations (present study) instead of two (Impact Study) raised the rank for engineering from 4.5 to 3; and interchanged social science/humanities and nursing/agriculture. The divisions are

not ranked so much in the same way, however, for cumulative grade point average and upper-class grade point averages (Rho= .42). Grade averages for physical science, biological science, engineering and humanities are relatively lower in the last two years (not actually so, but in terms of relative position in the rank order); while grades for home economics, elementary education and physical education are relatively higher. The rank difference correlation between tests and cumulative four-year grade point average is .46; the correlation with upper-class grade rank is -.39. (This difference is significant at the .01 level of confidence.) The point of this is that grades earned in the last two years (mainly in major field courses) serve to increase the disparity between grade-ranking and test-ranking, making it even more likely that to confer academic honors solely on the basis of grades will overlook some students who would be judged as equally or more capable by the external examinations.

Grades, Tests, and Temperament

"Conforming students tend to make higher grades, while independent, creative students of the same or higher ability make lower grades." This conclusion of a faculty subcommittee (above, p. 14) is supported by evidence from testing programs conducted by the University Impact Study and by the Student Counseling Service.

College Student Questionnaire (The University Impact Study)

The senior class of 1967 completed the College Student Questionnaire, a nationally standardized form constructed and scored by the Educational Testing Service of Princeton, N. J. The results were analyzed and reported by the Impact Study in two reports of limited circulation (C.F. Pemberton, 5/5/69 and C.F. Pemberton and A. Zawacki, 5/9/69).

The CSQ yielded student self-ratings in the categories: family independence, peer independence, liberalism, social conscience, and cultural sophistication. Also obtained were ratings of satisfaction with students, faculty, administration and major field, as well as ratings for study habits and extra-curricular participation. Average scores for 11 curriculum groups were reported, so that rank orders of such averages can be obtained, and comparisons made with "test" and "grade" ranks as in Table 5. This interpretation of the CSQ data is shown in Tables 7 and 8.

Table 7 shows that ranking of divisions by external examinations is positively correlated with CSQ rankings for Family Independence and Peer Independence, while the ranking by grades is negatively correlated with rankings for these two scales. The opposite kind of relationship occurs with Social Conscience, which is positively correlated with grade ranking. The dichotomy here is preference for autonomy and independence (associated with higher test competence) versus preference for structured and conforming social relationships (associated with higher grade competence). No significant difference was found between these two "types" with respect to Liberalism and Cultural Sophistication, although the coefficients were higher with test rank in each case.

Table 7. Correlations between Ranks on CSQ Group Scales and Ranks for Tests and Grades*

Academic Division	Test Rank	GPA Rank	Family Indep.	Peer Indep.	Liber-alism	Social Consc.	Cultural Sophis.
Phy Sci	1	6	1	3	4	8	6
Bio Sci	2	5	6	7	3	7	4
Engin'g	3	10	4	2	9	11	8
Soc Sci	4	9	5	4	2	6	2
Humanit	5	7	2	1	1	2	1
Home Ec.	6	2	7	10	7	4	5
Bus/Econ	7	11	3	6	10	10	10
Nursing	8	3	8	11	8	5	7
Agricul	9	8	11	5	11	9	11
Elem Ed	10	4	10	9	6	3	3
Phys Ed	11	1	9	8	5	1	9
Rho for CSQ Rank:							
with Test Rank			.76	.55	.40	-.46	.35
with GPA Rank			-.48	-.70	.21	.75	.14
Signif. of Difference between correlations			.001	.001	n.s.	.001	n.s.

A further analysis of the 1967 CSQ data is shown in Table 8. This tabulation shows that differences in attitudes toward faculty, administration, students, and major field are reflected in the ranking for grades and tests. Students in curricula ranked relatively higher by tests expressed more satisfaction with the faculty, with other students, and with their own major field; but less satisfaction with the college administration. The opposite relationship is observed for students in curricula typified by higher grades than test scores.

Students in curricula ranked highest for grade point average were more likely to question the validity of grades and grading procedures, as well as the competency of their instructors. They more often agreed (or failed to disagree) with statements implying that grades are based on irrelevant and extraneous factors, that students use "pull" and "bluff" to get through courses, and that their major departments reward conformity. They also

*Test rank and GPA rank in Table 7 are for the class of 1969; CSQ ranks for the class of 1967. While students might very well differ in the two classes on total scores, we probably can assume that department ranks would be essentially the same. The department rank orders for GRE (1967) and general test average (1969), for example, are correlated .90.

Table 8. Relationships Between Grade/Test Competence and Student Satisfaction with College Experience
(Rank difference correlations for curriculum group rank orders)

CSQ-2 Scales and Selected Items (From p. 11-13 of referenced report)	CSQ Correlation with: Test-Rank	Grade-Rank	Level of Signif.*
Satisfaction with Faculty. (esteem for instructors as competent, fair, and accessible)	.04	-.35	n.s.
a. Most are superior teachers	.36	-.89	.001
b. Most are competent in field	.39	-.46	.04
c. Welcome student disagreement	.36	-.37	.06
d. Grades not based on irrelevant, extraneous factors	.79	-.53	.001
Satisfaction with Administration (agreeable and uncritical attitude toward college administration)	-.50	.70	.001
a. Most college rules are necessary	-.18	.42	.10
b. Adequate help with educational plans	-.64	.59	.002
c. Rules are enforced impartially	-.77	.50	.001
d. Students not treated as children	-.66	.33	.01
Satisfaction with Major (positive attitude toward major department)	.58	-.07	.07
a. Dept. does not reward conformity	.45	-.30	.06
b. Satisfied with major field courses	.61	-.22	.03
c. Prestige of dept. relatively high	.75	-.38	.003
d. Satisfied with own grade standing	-.55	.89	.001
Satisfaction with Students (approval of student behavior and attitudes)	.74	-.19	.01
a. Not too extremist in politics	.79	-.23	.005
b. Satisfied with academic honesty	.49	-.36	.03
c. Not too nonconformist	.58	-.22	.04
d. Don't use pull or bluff to pass courses	.86	-.48	.001
Study Habits (serious and disciplined orientation toward academic obligations)	.01	.33	n.s.
a. Usually was prepared for exams	-.54	.68	.002
b. Kept assignments up to date	-.37	.50	.03
Extra-Curricular Involvement (amount and type of extra-curricular activity)	.18	.22	n.s.
a. Student government activities	.64	-.14	.04
b. Pre-professional organizations	-.56	.38	.02

*Procedure for testing significance of difference between correlations derived from: Guilford, J.P. Fundamental Statistics in Psychology and Education. New York: McGraw-Hill, 1965 Ed., p. 189-190.

expressed less satisfaction with the degree of academic honesty they saw around them. Students in those departments with highest grade averages, while satisfied with their own academic standing, were less satisfied with their major field course work and with the academic prestige of their departments.

The generally conforming temperament associated with GPA rank is shown (in Table 8) in the greater extent to which those students ranked higher by grades disapproved of political extremism and nonconformity among their classmates, while more often approving administrative rules and policies. (This does not imply that either viewpoint is better than the other, but simply points up the temperamental differences associated with test-taking and grade-making behavior.)

The rankings by tests and by grades did not differentiate students (or, to be precise, academic divisions which are composed of students) on the total scales for study habits and extracurricular involvement, although some individual items did so. For example, students in departments ranking higher for grades apparently followed a more unvarying schedule, reporting that they were usually well prepared for class examinations and kept their assignments up to date; students ranked higher by tests were more often involved in student government affairs, while those ranked higher by grades were more active in pre-professional, vocationally oriented activities.

The Freshman Testing Program (Student Counseling Service)

From evidence cited above, it appears that there are distinctive temperamental correlates of grade-making and test-taking behavior. Thus far, these relationships have been inferred from rank orders of average scores for academic divisions. A more direct comparison can be made from freshman test data.

The freshman testing program conducted by the Student Counseling Service obtains inventory measures on various aspects of interest, motivation, and learning style. Scores are interpreted to students but are not furnished to teaching faculty or to other persons concerned with the awarding of grades and academic honors. The inventory scores and grades are, therefore, experimentally independent measures. Five relevant measures are available for the class entering in 1965:

1. Achiever Personality (AP). Attitudes associated with grade-making, involving organized and conscientious application to assigned tasks, and affinity for the more structured aspects of learning; hence, "academic conformity."

2. Survey of Study Habits and Attitudes (SSHA). Related to the above--a measure of study methods, motivation for studying, and attitudes toward classroom academic activity.

3. Intellectual Quality (IQ). Attitudes associated with an "intellectual" or "theoretical" orientation to college, as opposed to "vocational" or "practical."

4. Critical Thinking Appraisal (CT). Mainly a test of formal logic, measuring attitudes of inquiry and respect for evidence as well as competence in inference and generalization.

5. Creative Personality (CP). A measure of some attitudes associated with creative temperament and behavior--imagination, originality, and nonconformity.*

Four external examinations were taken by the class entering in 1965: (a) the College Board SAT, (b) the College Level Examinations (CLEP), (c) the Graduate Record Area Tests, and (d) the Graduate Record Advanced Tests. These examinations are experimentally independent of teacher-assigned grades and of the five measures of temperament and learning styles, although we shall expect to find significant correlations.

From what has gone before, we could hypothesize that ~~(a)~~ college and high school grades would more significantly be related to AP and SSHA, while the external examinations would be more closely related to IQ, CT, and CP. This hypothesis can be tested by comparing mean scores on the temperament measures for "high-test" and "high-grade" students. Three student groups were identified from the class entering in 1965:

A. Honors--High Tests. Students graduated with honors (including high honors, honors, and distinction) either in 1968, 1969, or 1970, who also achieved an average composite score on the four external examinations of 600 or above.

B. Honors--Low Tests. Students graduated with honors or distinction who averaged below 600 on the examinations.

C. High Tests--No Honors. Students not graduated with honors but scoring above 600 on the examinations.

This sort of grouping selects out about the top 10% of the graduating class in terms of grades (those graduated with honors) and the top 10% in terms of external test scores. These categories are not mutually exclusive. The groups to be compared for temperament and learning styles are either (a) high on tests (600 or above), or (b) high on grades (3.25 or above), or (c) high on both. The comparisons are summarized in Table 10, showing mean differences between groups.

Three patterns or clusters of relationships show up in this kind of tabulation, in terms of the amount of difference between means and the algebraic sign of the difference: (a) variables related to grades, (b) variables related to tests, and (c) creative personality.

The Honors groups (high tests, low tests, and total honors) had significantly higher mean scores for college grade point average, as would be expected since that is the primary basis for awarding honors. Beyond this,

*For a more complete description of these tests and inventories, and evidence for their validity, see The Freshman Testing Program, (Pemberton, W.A. and Simons, E.N., Student Counseling Service, 1970).

Table 10. Mean Differences Between Groups: Honors/High Tests, Honors/Low Tests, Total Honors, and High Tests/No Honors

Variables	(A) Honors HiTests	(B) Honors LoTests	(AB) Total Honors	(C) HiTests NoHonors	Difference between Means* A-B	A-C	B-C	AB-C
Mean Scores for "Grade" Variables								
H.S. Rank	68.6	66.4	67.5	61.7	ns	+6.9	+4.7	+5.8
Coll. GPA	3.51	3.43	3.47	2.66	ns	+.85	+.77	+.81
Ach. Pers.	45.8	44.8	45.3	42.2	ns	+3.6	+2.6	+3.1
Study Hab.	44.9	43.2	44.0	40.1	ns	+4.8	+3.1	+3.9
Mean Scores for "Test" Variables								
SAT	664	591	628	652	+73	ns	-61	-24
CLEP	646	572	613	634	+74	ns	-59	-21
GRE Area	639	530	584	629	+109	ns	-96	-42
GRE Adv.	666	593	628	650	+73	ns	-55	-22
Int.Qual.	55.0	51.3	53.2	56.2	+3.7	ns	-4.9	-3.0
Crit. Th.	78.2	71.0	74.8	78.3	+7.2	ns	-7.3	-3.5
Mean Scores for Creativity								
Creat.Per.	39.0	38.7	38.9	41.9	ns	-2.9	-3.2	-3.0
N	67	67	134	106				

*All differences indicated by + or - are significant above the .05 level of confidence.

the honors group was significantly higher than the high-test/no-honors group on high school rank, achiever personality, and study habits and attitudes.

The high-test/no honors group resembles the high-test/honors group in having significantly higher SAT, CLEP, and GRE scores, since that, again, was the basis for selection. But more than this, the low-test/honors group (about half those receiving honors) scored significantly below the two high-test groups on intellectual quality and on critical thinking. Both honors groups (high and low tests) were significantly lower than the high-test/no honors group on the measure for creative personality.

Clearly, then, the awarding of honors on the basis of grade point average selects out those students who have efficient and diligent study habits, and who learned to operate that way in high school. Just as clearly, such a restricted basis for awarding honors has overlooked some well-educated students (as evaluated by external examinations) whose grades were almost as high, and who possessed in higher degree the valuable human qualities of initiative and intellectual curiosity. Of those 106 students not graduated with honors who had high test scores, 24 (23%) were graduated with final grade point averages above 3.00, and 51 (48%) had cumulative averages above 2.75.

In addition to these conventional indices of good scholarship, the "high-test" students showed more evidence of broad intellectual interests and creative temperament. Honors graduates as a group scored below the no-honors/high-test students on a measure of creative personality. This reinforces evidence and opinion previously cited that extreme concentration on grade-making, such as may now be required for graduation with honors, tends to inhibit innovative and creative performance.

It is not so much a matter of concern that some students are honored who do not deserve it--probably no student who received honors in 1969 and 1970, and few students who have been graduated, could be considered really unworthy of that distinction. What should concern us is that some worthwhile students are not being recognized.

Although not a solution, at least a partial correction is provided by the external examining committee which interviews all candidates for high honors. A comparison was made for the 1969 and 1970 graduating classes between those candidates for high honors who were accepted and those rejected by this committee. Students passed for high honors had significantly higher scores than those rejected on the Graduate Record Examinations and on measures related to intellectual orientation and creative temperament. There were no significant differences between the two groups in grade point average and on a measure of study habits. Since the test scores were not available to the committee, it seems evident that these particular test differences were reflected in some relevant behavioral qualities which influenced the examining committee. (Pemberton,C.F. and Pemberton, W.A., April 1970; and Pemberton, C.F., June 1970)

The "cluster analysis" shown above in Table 10 indicates that students can be differentiated into three relatively distinct groups, which correspond to the three "learning styles" measured by the Opinion, Attitude and Interest Survey (OAIS): (a) Achiever Personality (related to grades); (b) Intellectual Quality (related to tests); and (c) Creative Personality, inadequately defined by conventional academic measures but more related to tests than to grades. A factor analysis of these interrelationships probably would yield three distinct factors for these three "learning styles."

It has become clear in this study that different students learn in different ways. The University should adopt imaginative and diversified ways of evaluating students and provide better ways for bright and innovative students to be recognized. This is not possible if we continue to reduce all judgments to the ubiquitous grade point average, fitting all students into one Procrustean mold.

A Two-Fold Criterion for Evaluation

Several faculty members, after reading the reports of the Impact Study mentioned above, have suggested that academic achievement should be evaluated in terms of a combination of GRE scores and grade point average. There is no doubt that such a dual criterion would provide a more equitable balance between the sexes and among the academic divisions. To show how this might work in selecting students to be graduated with honors, Table 11 has been prepared.

Table 11 is a correlation scatterplot which shows the pattern of relationships between cumulative grade point average and external test average (CLEP and GRE) for students entering in 1965. The 951 students tabulated comprise

Table 11. Distribution of Grade Point Averages and Composite Test Scores for Students Entering in 1965 who have been Graduated by 1970.

Cum. GPA	CLEP/GRE Average 350	375	400	425	450	475	500	525	550	575	600	625	650	675	700	725	750	Totals for GPA M	F	T
4.0																				
3.9											1	1	2	1				2	3	5
3.8										1	1	2			1			1	4	5
3.7							1			2	2	1						3	3	6
3.6					1				1	1	3	3	4	1	1	1		7	9	16
3.5							1	1	2	3	4	2	3	3		2		13	8	21
3.4								1	3	4	6	2	1			1		7	11	18
3.3					1		1	3	6	7	8	3	1	1		2		11	22	33
3.2								1	2	3	3	4	2	3	1			12	7	19
3.1				1		4	6	13	9	8	4	2	2					21	28	49
3.0			1	1	2	8	9	15	10	7	4	4	3					26	38	64
2.9				1	2	4	11	12	7	8	3	4	2	3				22	35	57
2.8		1	2	2	5	8	10	11	10	9	5	1	2		1			26	41	67
2.7				2	4	9	12	20	8	9	6	6	2	1				36	43	79
2.6			1	3	11	13	10	17	11	11	5	2	3	1				45	43	88
2.5		2	3	5	7	10	15	11	11	8	5	2						41	38	79
2.4			2	2	15	9	17	12	9	7	6	4	1		1			48	37	85
2.3			3	8	10	12	18	13	10	3	3	2						52	30	82
2.2			3	4	11	12	11	10	14	5	2		1					53	20	73
2.1			3	6	7	8	11	8	8	6		3	2				1	47	16	63
2.0			1	8	7	5	6	5	5	1	3	1						35	7	42
N																		508	443	951
Total M	2	8	28	38	48	68	80	70	51	41	31	23	10	4	5	1		508	53%	
Grad- F	1	11	15	45	54	71	73	56	52	33	18	8	4	1	1			443	47%	
uated T	3	19	43	83	102	139	153	126	103	74	49	31	14	5	6	1		951	100%	
Actual M						1	1	3	6	12	5	6	4	1	5			44	9%	
Honors F				2		2	4	9	12	13	9	5	2	1	1			60	14%	
Distribution T				2		3	5	12	18	25	14	11	6	2	6			104	11%	
Proposed M						1		3	6	13	8	11	9	3	5			59	12%	
Honors F							1	3	12	15	12	7	3	1	1			55	12%	
Distribution T						1	1	6	18	28	20	18	12	4	6			114	12%	
Change in M							-1			+1	+3	+5	+5	+2				+15	+3%	
Honors F				-2		-2	-3	-6		+2	+3	+2	+1					-5	-1%	
Distribution T				-2		-2	-4	-6		+3	+6	+7	+6	+2				+10	+1%	

Proposed Honors Level

Present Honors Level

first-admission entering freshmen who went through the freshman testing program, and as of August 1970 had been graduated or had completed requirements for graduation (about 65% of the original entering class).

As a matter of interest, the resulting correlation between tests and grades when this table is solved, is .50 for women; .47 for men; and .44 for the total group. The fact that the correlation for total students is lower than that for either sex probably results from some sex-related bimodality in the distribution of grades and test scores. This further demonstrates that to lump all students together in terms of grade point average tends to confuse rather than to clarify the true relationships among students with respect to academic competence.

Table 11 may take some time to interpret, but there is no simpler way to show the relationships. Reading across the rows one finds the totals for grade point average, showing a higher proportion of women at the higher-GPA levels (above 3.00) and a larger number of men below 2.50. Reading down the columns yields totals for test scores, and here the relationship is reversed, with a higher proportion of men achieving test averages above 600.

The present system for conferring honors is represented by the horizontal line drawn above GPA level 3.2. All students above this line theoretically should have been graduated with honors. (Actually, three students above 3.25 did not receive honors; and three students below 3.25 were graduated with honors. This probably occurred because the final cumulative average for these students differed from that at the end of seven semesters.) Two women students were graduated with honors having test averages of only 450, and three honors graduates (two women and one man) had test averages of only 500. At the other extreme, 12 students (10 men and 2 women) had test averages of 700 or above; eight of these were graduated with honors and four (all men) were not. The highest test average (750) in this group was posted by a male student whose cumulative GPA after five years was 2.10.

Table 11 shows an alternative scheme for selecting honors graduates in terms of a dual criterion. The diagonal line drawn from test average 400 down and across to GPA 2.6 establishes hypothetical minima for these two indices of competence. All above this diagonal would be considered honors graduates. One might conceivably achieve honors rank with test average of only 400 and grade point average of 4.00; or with grade point average of only 2.50, paired with a test average of 800. It is unlikely that either of these extremes will actually occur. In Table 11 the extremes are: GPA 3.7--Test Average 500, and GPA 2.80--Test Average 700.

One consequence of such a procedure for awarding honors would be a more equitable sex balance. In this group 14% of the women were graduated with honors and only 9% of the men. With the new procedure there would have been 12% men and 12% women. There would have been a net increase of 15 men and a net decrease of five women in the honors group, as revised.

A similar plan could be used for setting graduation requirements, giving tests and grade point average appropriate weights. The details of such weighting easily could be worked out, provided the principle of external examinations and credit by examination is accepted. It might be possible, for example, that a student with grade point average of 1.75 and external examination average of 650 could be graduated. Who is to say that such a student is not as well

educated as the three students shown in Table 11 who were graduated with grade point averages above 2.50 and test averages below 400?

Table 11 does not tell the full story with respect to discrepancy between grades and tested competence. To my personal knowledge, four Merit Scholars whose CLEP and SAT scores averaged 700 or above are not shown at all in Table 11, having dropped out or been dropped by the University. Numerous other students (mainly men) with CLEP and SAT averages above 600 have withdrawn from the University, some failing and some passing. Of 30 students admitted to the University with Advanced Placement in 1965, five were graduated with honors and 25 were not.[1] There is no guarantee, of course, that a more flexible evaluation policy would have changed the situation for these students. It is likely, however, that some of these high-ability, low-achieving nonconformists would now be in graduate school if they had been able to obtain credit by examination, or if there had been some other alternative to the traditional evaluation procedures. Among these students may be some of the most creative minds in the entire class.

[1]Virtually all Advanced Placement students had test averages of 575 or above. Of the 282 graduates with test averages of 575 and above, 82 or 29% were graduated with honors. Among Advanced Placement students, only 5 of 30, or 17%, were graduated with honors. The Advanced Placement students apparently were penalized in terms of grades and honors by having begun their freshman year at advanced levels, or by being the kinds of people who would choose to do so. Grade-point-determined honors, it appears, go to the cautious.

CONCLUSIONS

At this point one must conclude that the University of Delaware, in common probably with most other institutions, and without malice aforethought, has been systematically discriminating against students who are: (a) male, (b) enrolled in the sciences and traditional "academic" disciplines, and (c) academic non-conformists. It seems beyond dispute that to evaluate students solely in terms of grade point average will obscure important differences in individual temperament, character, and general culture.

Grading has been examined and found wanting on three grounds: ethical, rational, and pragmatic. In particular, the grading system fails to recognize the creative and self-directed student. This, indeed, may be the basic issue. To say that women make better grades because they are better grade-makers is mere tautology; and it is pointless to debate whether departmental differences in grade averages are due to easier grading or to better teaching. Overriding the categories of sex and curriculum is the general dimension of conformity. Students differ in the degree to which they prefer structure, order, and well-defined tasks. Students who prefer order and dislike ambiguity are more likely to conform to grade-making requirements, and to gravitate toward programs of study which are highly structured in content and in teaching methods.

One of the most persistent dichotomies in metaphysics is Self-Other, whether termed introversion-extroversion, reflective-impulsive, conceptual-perceptual, innerdirected-outerdirected, or the Ying-Yang of Oriental philosophy. Difference in self-other orientation can be observed even among infants, and in students is a stable and predictable temperamental variable which influences individual learning styles. Some prefer to be taught and graded by teachers; other prefer to operate more flexibly, in their own way.

Most faculty members probably will agree with Professor Nagel of Berkeley that it would be deplorable if the rigorous and critical educational environment should give way to "a congenial, unevaluative one in which scholars went about their business and students were simply welcome to pick up what they liked as spectators on the intellectual scene." (Above, p. 10) Although evaluation solely by grades may reward the conformist, evaluation entirely based on external examinations might tend to reward that student, more clever than wise, who is adept at test-taking but unable to impose self-discipline and unwilling to accept external restraints and requirements. A two-fold basis for evaluation seems preferable.

It is plain that even if grading is not, as some maintain, wholly a capricious and arbitrary procedure which contaminates education, neither is the system as it exists adequate for this university and for these times. There is no simple remedy, but the problem exists and the problem is ours. We have to go along with Walt Kelly's Pogo who confessed: "We have met the enemy and they is us."

It does not get us out of the bind to say that the cumulative grade point average represents a consensus of different instructors, who should be able to evaluate student competence in their own fields. The fault is not in the teachers, but in the system. The fact remains that there are, as you have seen, systematic biases in the grading system, here at this University, which continue to operate in 1970, and cannot be explained away. The snark really is a boojum.

SOME POSSIBLE ALTERNATIVES

It is beyond the legitimate scope of this paper to recommend faculty action. The facts, as I understand them, have been presented fairly. Solutions should be recommended by a properly appointed faculty committee, but some review of what others are doing is in order. In his overview of student assessment procedures, Richards gives approving attention to four innovations: (a) pass-fail grading, (b) awarding of credit by examination, (c) criterion-referenced teaching and evaluation, and (d) procedures for evaluating creative extracurricular achievements.

Pass-fail grading. The widespread adoption of pass-fail grading has produced conflicting reports of success and failure. Students themselves seem ambivalent about the practice, partly I gather, because the grade point average still remains the crucial basis for evaluation, and a course taken and passed seems wasted for such computation. Nevertheless, this procedure solves some problems for some students, and seems likely to be continued and expanded.

Credit by examination. The principle of awarding course credit by examination, although an official policy of the University, is a procedure seldom used and little known to students. The difficulty in preparing individual comprehensive examinations is one of the drawbacks. There is no question that people can and do learn in a variety of ways, and that there is little point to repeating in a college course what one already knows, simply to obtain credit. The College Level Examinations (see above, p. 15-17) seem well designed for this purpose, and a growing number of colleges give credit by these examinations. The College Level Examination Board reports, for example, that in 1969 the University of Nebraska at Omaha graduated 800 students who averaged 20 semester hours of credit by examination on the basis of College Level Examinations.

As in pass-fail grading, there is some problem in equating credits earned in this manner to the yardstick of a grade point average. The goal of providing more flexibility in evaluation is a sound one which recognizes the validity of different learning styles and differences in formal educational backgrounds. The purpose of credit by examination is to recognize competence, however it is gained, even when not blessed by orthodox ritual forms.

Criterion-referenced evaluation. The idea of criterion-referenced testing has caught on since 1962 to such a degree that Richards refers to it as "a revolution in testing." The 1970 Annual Educational Conference sponsored by the Educational Records Bureau has announced its theme as "Testing in Turmoil," with special attention being given to developments in mastery testing or criterion-referenced testing.

During the summer of 1970, 15 faculty research grants were awarded to University of Delaware faculty members. The majority of the projects were concerned with the development of new teaching and evaluation procedures, some of which would represent moves toward the criterion-referenced model. Professors Neisworth and Crouse in Education, Cicala and McLaughlin in Psychology, and Markell in Accounting have tried what are variously described as "computer-assisted," "quasi-programmed," and "criterion-referenced" innovations in their classes. They report general student acceptance and significant improvement in achievement as compared with control classes.

The term "criterion" or "mastery" as employed in this context is used in the sense of standard for student performance rather than a variable to be predicted. A course is defined in terms of specific knowledge or attitudes to be learned, and a pool of test items is constructed to measure these objectives. The purpose of a criterion-referenced examination is to determine absolute level of mastery, not relative competence as compared with other students. The grade is in terms of pass-fail for each unit, although the final evaluation usually is determined by comprehensive examination. Each student proceeds at his own pace, as rapidly as he can master the content of individual units.

A key aspect of these methods is the use of student tutors who check the tests and review with each student the items missed, at that time. The tests are used for teaching and diagnosis, not for evaluation and grading. Reports from persons using such methods indicate that students usually learn more, in less time, and with greater satisfaction. Richards comments on the basic significance of criterion-referenced tests in this way:

> ...we now have a technique...that will provide information about the specific content mastered by each student without reference to the performance of other pupils...Because it would no longer be possible for a student taking the same examination to receive grades ranging from A to F depending on how bright the other students in his class were, competition for grades would be eliminated. This advantage... is not minor, for current grading practices almost universally treat courses as...competitive races (which) seems quite destructive of the values and goals of higher education. (Richards, 1970)

Specifications are being worked out by various persons for writing "perfect" criterion-referenced test items for a particular course, which involve precise rules for writing "distractor" alternatives as well as correct responses for multiple-choice questions. To those who worry that such procedures are "mechanistic," Richards replies:

> ...instructors are not really required to be mechanistic to write such tests. Rather, they are required to be explicit about the purposes of their courses--a requirement that should be damaging to few courses. Moreover, if an instructor thought he could not specify any skills or knowledge that students should have as a consequence of taking his course, it is difficult to see how he could justify assigning grades on any basis.

Creative achievement. A question that has long concerned teachers is how to identify and evaluate originality or creative performance. One scale of the OAIS (creative personality) was shown above (p. 27-31) to distinguish high-test from low-test students, and students graduated with high honors from those nominated for high honors but turned down by the examining committee. Such a measure of creative temperament, of course, does not necessarily identify students with actual creative accomplishments.

Richards believes that the College Achievement Scales developed by Holland for the National Merit Corporation provide "socially relevant measures which can serve as fairly comprehensive criteria of success in college," beyond grades. These measures of nonacademic accomplishment are only moderately correlated with

grades and scholastic aptitude tests, and their relationship to later adult accomplishment has not been studied. The theory behind the use of such non-academic measures is in line with Richards' conviction that "the single-minded pursuit of grades is destructive of other, perhaps more important values."

The external examining committee which interviews candidates for high honors at this University does succeed in identifying some of our more creative graduates, but only those who have achieved grade averages of 3.50 or above. There may be significant creative accomplishments from other students, not so high in grades, that are not recognized in any way but should be. If a direct assessment of creative accomplishment is not available, the broadening of the evaluation base to include pass-fail options, credit by examination, and external comprehensive examinations should in itself result in fewer creative students being overlooked.

Evaluation in the University Community Design

A faculty-student task force, chaired by Dr. Barbara Settles, has been appointed by Vice-President John W. Shirley and instructed to study in depth the issue of student evaluation. The following is an effort to get discussion started and elicit reactions from the faculty, but does not constitute a consensus or recommendation of the task force.

1. Evaluation of applicants for admission to the University should be based on the likelihood of their eventual graduation rather than the probability of their achieving a minimally acceptable grade point average during the first year. Rather than a PGI (predicted freshman grade point index) one could compute the PG (probability of graduation) for that student. Admission testing would be more in the nature of placement testing, concerned with where a student should start and not merely with whether he should.

2. A General Studies Division would include College Try students and undeclared "students-at-large" not yet admitted to a particular college. Special tutorial, remedial, and counseling services, as well as flexible scheduling strategies, would be employed with these students until they become admissible to a college or become clear enough about their purposes to choose a particular college. Some might be persuaded to choose a different kind of institution altogether. The General Studies Division would work closely with the Student Counseling Service, and counseling psychologists might act as faculty advisers to these students until such time as they are admitted to an academic college.

3. Students would be admitted to a general program in each college, declaration of a major not being required although not forbidden. Fewer courses would be pursued at a time, but at a more intensive level. Classes would be conducted with a view to helping all, or a substantial majority of students to achieve "mastery level," via criterion-centered rather than norm-centered teaching and evaluation. The blame for failure would be shared by teacher and student.

4. When an adequate level of competence (determined by courses passed or credits earned by examination) has been demonstrated, a student could apply for admission to candidacy for the degree--normally at the end of the sophomore year, but sometimes earlier or later. The terms "freshman," "sophomore," etc., would gradually come to have less relevance in such context. There would be three

main categories of undergraduate students: (1) General Students in the General Studies Division, (2) underclassmen, and (3) upperclassmen, the latter comprising those who have been formally admitted to degree candidacy in a particular department. Those not admitted to candidacy could settle for an Associate Degree.

Admission to candidacy (or upperclass status) would be determined by (a) level of general culture as determined by competent examinations such as the CLEP general examinations, (b) performance on a comprehensive examination in the proposed field of study, and (c) underclass grade point average (with courses, if any, in the General Studies Division not computed in the GPA although counted as courses passed for credit toward admission to degree candidacy).

Once admitted to upperclass status the computation of grade points for each semester would be discontinued as a requirement for staying in college, although grade point average would still be computed and reported for the record, and the cumulative GPA would be one criterion for graduation. Some tangible evidence of "progress toward the degree" would be required for continuation as an upperclass student, but the pressure to achieve grade points would be reduced. The admission-to-candidacy requirement should be rigorous enough that an upperclass student would have virtual assurance of being graduated, provided of course that he does not change his mind about being a student or fall apart in the meantime.

5. Graduation would be determined by a two-fold criterion, grades and external examinations, and graduation with honors, or at least with high honors, would require additional evidence of independent or creative accomplishment. Graduation requirements would include (a) general competency as shown by the Undergraduate Record Area Tests, (b) special competence in major field as shown by the UGRE Field Tests or equivalent, and (c) upperclass grade point average. The relative weights given to these various criteria would probably vary from college to college, but certain minima would be required in all degree programs. A large discrepancy between grade point average and external examinations would constitute grounds for review and perhaps reexamination.

6. The implementation of such a plan would require an Office of Examinations, which would be commissioned to help with the construction of criterion-referenced class examinations, administer all external examinations, and prepare comprehensive examinations for those colleges and departments for which appropriate standard measures are not available. The Office of Examinations would need to be relatively independent of the instructional departments, assisting teachers to teach but relieving teachers from the sole responsibility for evaluation.

While this may seem visionary it could be that the University of Delaware, at this time in a process of self-examination and open to innovation, is the very place to try it. We could become "the very model of a modern major university."

BIBLIOGRAPHY

1. Berkeley. Education at Berkeley. Report of the Select Committee on Education to the Academic Senate. University of California at Berkeley, March 1966, 93-103.

2. Boozer, H.R. External examinations as predictors of competence. Journal of Teacher Education, 16, 2, 1965, 210-13.

3. Bruner, J.S. The process of education. Cambridge, Massachusetts: Harvard University Press, 1960.

4. Bruner, J.S. Toward a theory of instruction. Cambridge, Massachusetts: Harvard University Press, 1966.

5. Dressel, Paul. The role of external testing programs in education. Educational Record, Spring 1964, 161-66.

6. Educational Testing Service. Multiple-choice questions: A close look. Princeton, New Jersey, 1963.

7. Geis, Florence. Machiavellianism in interpersonal relations. Paper given at Delaware Psychological Association, Annual Research Day, May 9, 1969.

8. Goodman, Paul. In what ways does the present marking and credit system inhibit or promote learning? Proceedings of the 19th Annual Conference on Higher Education, April 19-22, 1964.

9. Hewitt, R.G. The status of pass-fail options at 22 colleges and universities. Office of Institutional Studies. Amherst: University of Massachusetts. March 21, 1967.

10. Hoyt, D.P. The relationship between college grades and adult achievement. ACT Research Reports, No. 7. Iowa City, Iowa: American College Testing Program, Sept. 1965.

11. Lannholm, G.V., Marco, G.L., and Schrader, W.B. Cooperative studies of predicting graduate school success. Graduate Record Examinations, Special Report No. 68-3. Princeton, New Jersey: Educational Testing Service, 1968.

12. Mayo, S.T. Mastery learning and mastery testing. National Council of Measurement in Education, Special Report. Vol. 1, No. 3, March 1970.

13. Phi Beta Kappa. Report of pass-fail study committee to the senate (Karlem Riess, Chairman). December 6, 1969.

14. Pemberton, C.F. An evaluation of the cumulative grade point average as a means of identifying superior students. University Impact Study. Newark: University of Delaware, 1966.

15. Pemberton, C.F. A comparison of high ability under-achievers with low ability over-achievers. University Impact Study. Newark: University of Delaware, March 1969.

16. Pemberton, C.F. The student-functioning scales of CSQ-2 analyzed in terms of curriculum group. University Impact Study. Newark: University of Delaware, May 5, 1969.

17. Pemberton, C.F. and Zawacki, A. Change in scores between the freshman and senior years on the five scales common to CSQ-1 and CSQ-2. University Impact Study. Newark: University of Delaware, March 1970.

18. Pemberton, C.F. The relationship between grades and two external measures of academic achievement. University Impact Study. Newark: University of Delaware, March 1970.

19. Pemberton, C.F. and Pemberton, W.A. Memorandum to the Faculty Senate, concerning the degree with high honors. April 1970.

20. Pemberton, C.F. Comparison of 1970 high honors graduates with two eligible groups rejected for high honors. University Impact Study. Newark: University of Delaware, June 1970.

21. Pemberton, W.A. Ability, values, and college achievement. Newark: University of Delaware Press, 1963.

22. Pemberton, W.A. and Simons, E.N. The freshman testing program. Student Counseling Service. Newark: University of Delaware, 1970.

23. Reitz, H.J. Career orientation and academic achievement among elementary education majors. Counseling Psychology, 1970, 3, 205-209.

24. Richards, J.M. Assessing student performance in college. Eric Clearinghouse on Higher Education, Report 2. George Washington University, Washington, D.C., May 1970.

25. Riesman, David. Where is the college generation headed? Atlantic Monthly, 1961, 4, 39-45.

26. Rosenthal, R. and Jacobson, L. Pygmalion in the classroom. New York: Holt, Rinehart and Winston, 1968.

27. Rosenzweig, M.R. Environmental complexity, cerebral change, and behavior. American Psychologist, 1966, 21, 321-332.

28. Stoddard, G.D. The meaning of intelligence. New York: Macmillan, 1943.

29. Tennessee. Teaching-learning issues. Learning Resources Center, Report No. 2. Knoxville: University of Tennessee, 1966.