Grades: An Evaluation of Evaluations

Dierdre Kelly

October, 10, 2010

Everyone in this room has received them, mainly the good variety. Many people in this room have given them. Grades. Numerical or alphabetical, they start judging your performance in middle school, and continue for the rest of your formal education. Some disclaimers are necessary before I begin. I would like to immediately state that the purpose of this paper is not to attack those who give grades. I am not an expert on grading, and I merely use the information provided by experts in peer-reviewed papers and books. This paper is an invitation for everyone to examine the weight we as students and evaluators of students place on grades and the GPA, and consider the resulting consequences. Throughout the paper I will refer to both cited examples and familiar anecdotes that illustrate each point.

We’ve all heard it before, “He has a four-point-oh, he’s really smart.” Implicit in this statement is the equation of GPA and intelligence. Other qualities we attribute to someone with a high GPA: capability, creativity, diligence, and intellectualism. Do grades really reflect these qualities? Grades actually reflect numerous other qualities and events in a student’s life. However, graduate schools continue to use GPA as an item to compare and judge students. The issue is that graduate schools and related entities that use GPA to judge candidates assume that GPA represents desired qualities in the applicant. However, before delving into what grades reflect, we must first look at grading systems and how they developed.
Grades were first given in the United States at Yale University in 1783. The scale consisted of *optime* (honor men), *second optime* (pass men), *inferiores* (charity passes), and *peiores* (unmentionables). Yale began giving grades on a scale of 1 to 4 in 1800. The Grade Point Average was then born. Harvard began using letter grades in 1886, which were associated with percentages for grading. However, the grading wasn’t standardized, while Harvard’s failing grade was a 26%, Mount Holyoke set theirs at 75%. In the early 1900s, University of Missouri Professor Max Meyer noted that grading practices varied wildly among the school. He proposed that grading should be based on a distribution curve in which the top 3% would be judged *excellent* (*A*); the next 22%, *superior* (*B*); the middle 50%, *medium* (*C*); the next lowest 22%, *inferior* (*D*); and the bottom 3%, *failure* (*F*). The pass-fail system, while emerging in the mid-1800’s without much success, reappeared in the 1960’s with vigor. However, students and professors lost enthusiasm for the system and it quickly became simply an option. Since the 1960’s most schools have also added plusses and minuses to grades for increased assessment precision.

Eiszler outlines several purposes of grades in his 1983 taxonomy. They:

“(a) allow agencies and institutions to make discriminations among individuals and their performance, (b) motivate learners, (c) give information to learners about the quality of their performance, (d) give limited information to the teacher about the quality of instruction, and (e) meet a variety of administrative and institutional needs related to the functioning of the institution.” Not one of these objectives indicates that grades should serve as a reflection of learning, intelligence, diligence, or other desired student traits. However, the purpose (c) regarding quality of performance comes close. How purpose (c) informs the other purposes is of interest, especially how it informs purpose (a), or the
ability of grades to allow other agencies and institutions to discriminate among individuals. It is important to note here that grades serve multiple purposes, and in serving all purposes it may be less relevant to directly connect “quality of performance” with ability to be judged by an outside agency or institution. This is illustrated in the example of requiring the class average in a particular course to be a B-, which will be further discussed in a moment.

Grading systems and methods will first be examined.

Grading is intrinsically subjective. Every institution has a different system and different expectations for evaluating students. In addition, different departments vary, and to make the picture even more complicated, individual faculty members have entirely different grading systems, expectations, and rubrics (1). This often remains true even when different professors are teaching different sections of the same course. One particular grading system may only be based on the grades of a midterm and final exam. Another system may grade the weekly quizzes and student participation, and give final grades based on improvement over time rather than the average of all assignment grades. It is not unreasonable to say that a student’s performance in a course would earn an A from one professor and a B from another. This occurrence is noted by Professor James Hammons of the University of Arkansas, who found that, “the variation in grading practices among faculty within a given institution is so great that it cannot be said that an A is always an A or that any letter grade represents some definite achievement level. Consequently, it is a mistake for any institution to rely on grades in making discriminations among students.” However, it is unlikely that professors would standardize grading systems to each other, let alone departments and institutions.
One individual may also grade variably. The subjectivity in grading is not only related to a professor’s intended system, it may also be entwined with the professor’s attitude while grading, which is affected by time of day, location, illness/health, stress, and other factors. Subjectivity is likely to be more of a problem with grade assignment based on general feelings, i.e. “It feels like a B-.” rather than grades based on certain criteria as in the “Primary Trait Analysis” (PTA) method. When this method is used, grading becomes more similar over time, making grades standardized. It is easy for one professor to standardize their grading personally.

(2) An interesting dilemma arises with the national problem of grade inflation in post-secondary institutions. In the 1999 school year, 44% of final grades given at Dartmouth were A’s or A-‘s. Similar rates are occurred and are continuing to occur around the nation. Some instructors and administrators have commented on the fact that SAT scores of admitted students have risen, indicating better students, so college GPAs of these students should also rise. Harvey Mansfield, a professor of government at Harvard University, opposes this idea, stating that it is similar to evaluating your car based on the auto standards of 20 years ago. Yes, in that comparison, your vehicle will score exceptionally well, but it is not valid. In other words, students should face new higher standards of excellence. While grade inflation in particular, though an important issue, is not the focus of this paper, the variability in how colleges choose to deal with grade inflation affects the grading of students at those institutions. This is in addition to the fact that by nature, students at schools with grade inflation with have higher GPAs than they would at other institutions without grade inflation.
Let’s return to grading comparisons within institutions rather than between institutions. There are several forms of grading systems, each with their own pros and cons. It should be noted that all of them have shortcomings with regard to what the grades actually represent.

The first form is called Norm Referenced Grading, which grades students based on other students in the class, otherwise known as “curve” grading. This system allows separation of students into distinct groups of performance and results in a perfect average. However, it does not place the students in the context of the larger institution, in which they might be doing much better or much worse than the institutional average. Norm Referenced Grading does not measure learning, because it does not measure how proficient the student is with the material. The system only grades how proficient a student is to another student. In this way, very competent students may be given a final grade of a “D” compared to their peers. Because helping another student would change a student's own grade, this system also ruins cooperation between students that would aid their learning.

Criterion Referenced Grading is equivalent to the Primary Trait Analysis method previously referred to. In this method, a student’s performance is measured based on an assigned set of criteria. This system can technically measure what students have learned, if the criteria are properly designed. However, the selection of criteria is again subjective. The criteria may or may not be available to the students before they complete the assignment. In this manner, however, students are not graded on their mastery of the subject, but how well they fulfill a particular set of criteria. The criteria may not reflect mastery of the subject being graded. In addition, managing of grade distributions is impossible. Students may all do exceedingly well on an assignment, or they may all fail. Given the directive to
maintain a B- average of each class, professors would have to determine a way to achieve a B- average for the class by the end of the course. The main concern with this method is how to create and defend criterion levels for each resulting grade.

Another method of grading is based on “Mastery Learning;” a very generalized and perhaps vague method with variable applications. The premise of mastery learning is that students will be attaining total mastery over a given subject. The students will have access to appropriate individual instruction and sufficient time to achieve mastery. Frequent feedback is provided over the course of instruction, and students are graded on the level of mastery they have achieved. While many people would call this method the ideal model of learning in that students have every chance available to learn completely and thoroughly about a subject, this method is not feasible for a semester-based college course. Every student would be working on their own individual topic, there are no specific assignments to grade, and students will be at different places in their projects at different times, requiring intensive attention by the professor. Additionally, it is unlikely that students can achieve mastery in a subject in one semester. Lastly, because grading is based on an uncharacterized “mastery” value, grading may be vague, and many students may not have fulfilled this “mastery” value, resulting in a grade of “incomplete.”

The final method of grading is the familiar pass-fail system. This system generally refers to the fact that grades are assigned as usual in the course for assignments, but final grades may only be recorded by the registrar as pass/fail, satisfactory/unsatisfactory, etc. Advocates of this system hope that it encourages students to venture outside their normal course of study because of the lack of anxiety about their grades. Removing the anxiety and need to achieve may in fact result in greater learning. While pass/fail grades avoid
incorporation into GPA, pass/fail courses are not looked well upon by graduate schools, especially when a large number of them are taken. This is due to the fact that specifics regarding what the student has learned and how the student has achieved are unobtainable. Truly, the student may have taken the pass/fail course in order to do less work and aim for a just-passing grade. Additionally, students who do take pass/fail courses to explore a new area of study without anxiety may be in fact digging themselves a larger hole. These students are conditioned to receive grades as a reward and may suffer psychologically when grades are not given.

Most systems of grading can be traced back to these four categories. Each of these systems in some way returns to the same problem: grading does not reflect learning because it is aimed at measuring another quality as with class comparison in norm referenced grading, or avoiding another quality, as with anxiety in pass/fail grading. While some advocate a combination of these grading systems to amplify the benefits, unfortunately this combination also amplifies the detriments.

Up to this point we have established that grades serve multiple purposes that may be in conflict with each other, that grading is inherently subjective, is strongly affected by grade inflation, and can be categorized into systems that all have disadvantages. However, what do grades really represent? Why can’t the issues that classify what grades really represent be remedied? Why does it really matter that grades and GPA don’t really reflect a student’s capability? And what can be done about it if this is all true?

College life sets up several unique obstacles that must be overcome in order to do well. Students take four to five courses, and are usually involved in several additional activities such as athletics, employment, extracurricular clubs, and honor societies.
Students are often pressured to "get involved" on campus, and the choice to participate or not in these activities may not be a free one. This leads to a heavily scheduled life. A student may have class from 8am to 4:30pm, then athletic practice until 7pm, dinner, and employment at the library from 10pm to midnight. It can be said that a college student's career ranks high among those with the most scheduled time. This often leaves minimal time to complete out-of-class work, of which there is always a substantial amount. While every college student should be able to balance scheduled time and classwork, it is very easy for external factors to tip this balance. Three large assignments or exams in one week would undoubtedly result in lower grades on these items than if those assignments and exams were separated. Illness, as well as family emergencies, greatly affect school performance, especially if the severity of the occurrence is great enough to affect schoolwork but minor enough to not require a leave of absence. Ability to set oneself up for success in a course also depends in part the ability to afford the required resources for a course, be it books, audio and video resources, art materials, or molecular modeling sets.

These issues may seem simple to correct. Since every student is likely to have a bad week, one may reason, why not just drop the lowest quiz grade? If a student is ill (and has a note from Appel confirming this), allow the student more time on assignments, or let the student take an exam at a different time. If a student misses work due to leaving school to see an injured family member, let the student make up the work. Each of these practices is deemed “unfair” in several papers regarding the fairness of grading. Professor Daryl Close of Heidelberg University goes into great detail regarding the fairness of these practices. Dropping the lowest quiz grade of each student is inequitable because the quiz that is dropped is essentially random, in that it will be different for each student. Thus, students
are graded on different material. The grades of each student’s lowest quiz grade will differ. Some students may skip class, miss a pop quiz and receive a zero, others may attempt the quiz and receive a 70%, less than they would prefer. Is it justifiable for the professor to allow each of these grades to be dropped? The average of the rest of the student’s quiz grades is changed by not including this lowest quiz grade. The fact that all students are “wronged equally” does not justify this practice. Heidelberg finds that in order for grading to be impartial and consistent, if make-up work is given to one student because of illness or family emergency, then that make-up work must be available to all students should work be missed. It should be considered what this means for the student who never misses any work. If one aims to be totally impartial, then if additional time is given to one student, additional time should be given to all students. This seems to abolish the reason for giving the ill student additional time in the first place. The additional time would be a sort of handicap: a chance for the ill student to bring his or her grade up to the quality of the rest of the students. If the rest of the students are also given more time, their grades would supposedly also rise. This is less of a problem in criteria-based grading than in grading on the curve. The message here is that there would seem to be unfairness to students in both attempting and not attempting to assist a student having difficulties in a course. This unfairness is manifested in grading.

Each letter grade in an undergraduate’s career contributes to the overall cumulative GPA. After describing the complications that arise, is a 3.8 really different than a 4.0? Is a 3.5 different than a 3.8? Is a 3.0 different than a 3.5? Can you really say that a 3.0 is different than a 4.0? What I mean to say is, can you say that a student with a 3.0 doesn’t have the same capability, intelligence, or diligence as a student with a 4.0? The same
question applies for a GPA of 2.0 and a GPA of 3.0. One can’t really determine this unless the same student could be put through all the different permutations of classes and professors at an institution. Distinctions between 3.0 and 4.0 are being made. This is not shocking, although I find it unfortunate. However, the fact that distinctions are being made between a 3.6 and a 3.8, this is shocking.

Grades act as a form of academic currency that students can use to obtain honors, scholarships, and admissions to graduate programs. Numerous studies have found that GPA remains an important factor that graduate schools consider when admitting applicants. Some programs immediately weed out applications with GPAs lower than a specific number before any human reads them. The assumption is that the undergraduate GPA is an indicator of qualities that would make a good graduate, medical, law, business, etc. student and eventual professional in the respective careers. We’ve explored the complications of grading that eventually contribute to the final GPA. Admissions committees are not ignorant of the differences in grading between institutions, departments, and individual professors. Why then, do admissions committees continue to place so much weight on the GPA? The answer became clear during interviews with some admissions officers. Their answers varied on the following theme: “There are so many excellent candidates,” they said, “We need some way to discriminate and narrow down the applicants that we can look at for admission. Parsing apart candidates by GPA is a way we do that.” It is true that GPA is not the only factor in admissions decisions, however, I believe that using the GPA in this fashion is flawed.

I cannot tell you that undergraduate GPA never predicts performance in graduate school. However, the results of studies examining such correlations are mixed. A study
conducted at a “small Midwestern university” found that there was only a modest (read: little) correlation between undergraduate GPA and performance in the university’s graduate business programs. Other studies find strong correlation, as Michigan State University did when it looked at undergraduate GPA compared to performance in their criminal justice graduate program. The bigger question becomes, why is there variability? To that question I propose there is variability for the reasons described in this paper. Now, what can be done about it?

At first I believed that grades and GPA should be isolated from admissions processes, that grades and GPA should still be calculated for the benefit of the student for assessing their own performance in college, but never released to outside institutions or entities. Then I realized how unfeasible this was and that it would create utter chaos as these institutions would try to obtain GPA values anyway, perhaps by asking the student in the application, much like they ask for ethnicity. Professor Walvoord of Notre Dame and Professor Anderson of Towson wrote the following in their book Effective Grading: “We agree that grades, particularly final course grades, as isolated artifacts, are not particularly useful or appropriate for most contemporary institutional assessment needs. If the criteria, standards, and context for a grade are not known to external audiences, grades convey little information.”

While they did not directly propose a way to enact this sentiment, there is a way to do so. All final grades should be accompanied by a brief statement in the transcript. The paragraph should contain the grading system used (how many assignments, is it criteria based/norm referenced?), as well as the class average for the course. Any other information contributing to the student’s performance should be briefly described. Perhaps
the student excelled in the lab portion of the course. Perhaps the student had a severe illness that impacted the grade, but the professor still thought highly of the student. This differs from recommendations, in that not only are a few detailed assessments given to admissions committees, but every assessment that the student has ever received can be explained. This accompanying statement practice may even reduce the rampant grade-mongering disease among students, if their letter grade evaluation has context. Admissions committees would then have a more accurate picture of the student, with more sensitive assessments. The drawback here is the additional work on both professors for writing the statements and admissions committees for reading them. However, I believe this is attainable, since Dartmouth already includes median grades and additional information on final grades on their transcripts.

I hope to have complicated the discussion to proceed on grading this evening. I’d like to repeat that in no way am I attacking anyone with this paper. I merely wished to examine the ways in which grades are subjective, variable, and perhaps inconclusive as to the student qualities that we attribute to them. Perhaps these issues concerning grades would not be as important if grades were not used in scholarship and admissions decisions, but they still will be. Final grades accompanied by statements would give context for the evaluation, although it would require more time and resources. We have applicants with ever-increasing performance statistics entering undergraduates institutions. They deserve to have accurate and detailed evaluations. Thank you very much.

*For references, please contact Dierdre Kelly at dierdrek@gmail.com*