2002

Making the Grade: Some Principles of Comparative Grading

Jeffrey E. Stake
Indiana University Maurer School of Law, stake@indiana.edu

Making the Grade: Some Principles of Comparative Grading

Jeffrey Evans Stake

While I was a law school student, one teacher attempted to give grades of D or F to a large portion of his first-year class. Did he not know—I wondered—that it was statistically much more likely that his grading had changed than it was that the effort and ability of a class of 125 students was radically different from the effort and ability of his other recent first-year classes?[1] Upon joining the law teachers, I found that most colleagues were concerned about their grading but had spent little time thinking about how their grades fit into the context of marks given by their peers. They had not considered whether grading practices that would be benign standing alone could be problematic in that larger context. Most of us teach that process is important, but in some ways we do not practice what we preach. We spend little of our faculty energy overseeing our institutional process of calculating summary statistics such as grade point averages and class ranks.

This article is an attempt to bring important grading issues to the attention of law teachers. I hope to establish certain principles of comparative grading, what teachers should do in assigning grades to students and what schools should do in establishing grading scales and combining grades into grade point averages. I will set out, and offer some justification for, five grading principles—equalize means, equalize standard deviations, use numerous grade intervals, maintain proportional intervals, keep the no-credit grade reasonably close to the mean—and one corollary, avoid grade inflation. In part I, I mention briefly a few reasons that grading is worth our attention. In part II, I admit a few assumptions, some issues that will be left unexplored. In part III, the heart of the paper, I discuss the principles of grading set out above and

Jeffrey Evans Stake is a professor of law at Indiana University—Bloomington.

For helpful comments, I thank Pat Baude, Ken Dau-Schmidt, Terry Denny, Bob Heidt, Steve Heyman, Joe Hoffmann, Eric Rasmusen, Lauren Robel, Robert Stake, and Marjorie Young. Michael Alexeev, Mitu Gulati, and Russell Korobkin offered extensive and pointed written comments. John Applegate read it twice, something I would not ask of anyone. I thank Gerry Spann for encouraging me to turn my thoughts into this article. Articles in education journals have used "Making the Grade" in their titles, but my research assistant could not find that title in the law reviews.

1. Another possibility is that his students had learned less than earlier students because his teaching had changed. The question whether a terrible teacher should be allowed to give low grades is the flip side of whether a good teacher should be allowed to give high grades, an issue discussed below.

some possible exceptions to the principles. My brief conclusion is not intended to be the end, for I do not pretend that my thoughts will be the last word on any of these topics.[2] My goal in publishing this article is to start a discussion from which we can learn more about this weighty dimension of our responsibilities as teachers and faculties.

I. Why Do Grades Matter?

Before discussing the principles, some might want to take a step back and ask why we should bother. After all, grades say little. Grades generally purport to tell only who performed better and who performed worse on an instrument of assessment, usually a single exam or paper. To generate a bit of the *in terrorem* effect, many of us say that we will include class participation, but in most large classes it does not end up counting for much. Test results explain most of the variation in grades. The nearly ubiquitous grade point average is usually just a weighted average of all of a student's classes, a numerical combination of incommensurable grades measuring various dimensions of ability and learning. And class rank is normally derived from GPAs. Being based entirely on grades, these statistics add no new information to the thin account published in the individual grades.[3]

Although grades, GPAs, and class ranks contain specific messages that are quite limited, they are read to mean more. Our grades communicate broader meanings to the students and their potential employers. The signals we send to students relate to both whether they are studying well or enough, and whether they have the aptitude for a career in the law. Grades can influence the ways students think about themselves, swelling their heads or shaking their confidence. Student members of law journals use first-year grades to determine, to a large extent, which students will be invited to share the benefits of the law review experience.

Some believe that one of our most useful functions as law teachers is to sort students for their employers.[4] And whether or not it is our purpose to sort, grades do have that effect. Employers act as if grades reflect aptitude for being a lawyer. Many of the most prestigious and high-paying firms limit their hiring to students above a specified grade point average or class rank, or limit their interviews to students who are on a law journal, which is usually determined in part by grades. Grades and law review also determine who will have a chance to clerk for a judge, which is helpful both for developing legal skills and for

2. Nor do I mean to imply that mine are the only principles. For methods of adjusting GPAs, see Lawrence J. Stricker et al., Adjusting College Grade-Point Average Criteria for Variations in Grading Standards: A Comparison of Methods, 79 J. Applied Psychol. 178, 178 (1994); John W. Young, Grade Adjustment Methods, 63 Rev. Educ. Res. 151 (1993).

3. I have elsewhere advocated richer descriptions of student performance. Jeffrey E. Stake, Who's "Number One"?: Contriving Unidimensionality in Law School Grading, 68 Ind. L.J. 925 (1993). See also Michael E. Levine, Toward Descriptive Grading, 44 S. Cal. L. Rev. 696 (1971). Class rank does add useful information for a reader who does not know all of the other GPAs at a school, but it does not add information to that contained in all of the GPAs.

4. See generally Michael Spence, Market Signaling: Informational Transfer in Hiring and Related Screening Processes (Cambridge, Mass., 1974); Richard A. Ippolito, The Sorting Function: Evidence from Law School, 51 J. Legal Educ. 533 (2001).

gaining entrance to certain careers, such as law teaching. For our part, when we act as employers of new faculty, we law teachers spend little time reviewing the applications of C students. Grades affect which students win the honors and which are eliminated from some career tracks. We can argue about the degree to which grades matter, but few doubt that grades do matter.

Because law school grades send messages, messages that open and close doors, it is obviously unfair to say students performed differently when they in fact performed the same. It is also unfair to a higher-performing student to say that she performed the same as a student who performed much less well.[5] Of course two students rarely perform exactly the same, so there is always some unfairness in using the same grade twice, but that unfairness grows with the difference in their performances. Finally, and somewhat less obviously, it is unfair for a teacher to say that Sally performed "much better than" Jane, and Jane performed "better than" Dick, when in fact the difference between Dick and Jane was the same as the difference between Jane and Sally. Of course the same unfairness can occur in the context of GPAs and ranks as well as with individual grades.

In addition to being unfair, inaccurate communication via grades can also be inefficient because it misleads employers and it fails to set up appropriate incentives for students. A discouraged student may drop out of school. An unduly encouraged student may gain a false sense of confidence that he can do legal work without as much preparation as others. An employer may miss an opportunity to interview a student who would have been best for the job. Grades would hardly serve the sorting function if they were assigned randomly. At the extreme, inaccurate grading could lead firms to ignore our grades, just as many have learned to ignore our references.[6] Even if we do not care much about whether our grades facilitate efficient hiring, we should care if our grading practices drive frustrated employers to other schools for their new hires.

Although much of the discussion in this article relates to the goal of accurate communication, there is another important normative goal that is implicated by teachers' grading practices. Differences in grading methods create incentives that influence students in important educational choices. Students, and even their counselors, pay attention to the ways teachers grade; the grades teachers have given in the past are a factor when students pick their courses. For many students, it does not matter that most teachers would like students to exclude expected grades from their course-choosing calculus. Students also pay attention to our grading practices when they allocate their limited time for studying. Again, teachers would prefer that students not study more for one course simply because of the teacher's grading methods. The failure of teachers to adopt consistent grading practices influences student behavior in ways that teachers find undesirable.

5. It can also be unfair to a student to tell him that he performed better than he did. Students need to know when they did poorly, just as they need to know when they did well.

6. I have been listed as a reference on scores of résumés, but I can count on one hand the number of times I have been called by potential employers.

For all these reasons, it is our duty, to our students and their employers, to avoid miscommunication in the messages we send about how students performed. But how? Law teachers employ many different approaches to grading. Some of us add up points. Others use a gestalt method, holistic and sometimes intuitional, assigning a final letter grade without having assigned numerical scores at any stage of the process. Some teachers combine these methods by assigning initial points and then adjusting the final score in light of a complete rereading of the exam. Others take the opposite tack, assigning gestalt grades, but adjusting those up or down for particular points made or not. Many faculties include teachers from a number of these camps. However we generate the initial scores, does it help to draw lines in the gaps? Does it help for a school to prescribe a curve? Does grade inflation do any harm? The principles discussed here should apply to teachers in every camp and to all schools that either combine grades into grade point averages or publish grades that others will combine into grade point averages.

II. Assumptions

This article proceeds on a number of assumptions, many of which raise questions that are substantial enough to deserve articles of their own. In addition to those listed below, I am sure that other assumptions are important to the arguments I will make. But the list below should be sufficient to warn the reader of the type of question that I will leave unexamined.

1. Grades communicate relative performance.

I assume that grading communicates information about a student's performance relative to the performance of other students in the same school.[7] Not all teachers employ grades to this end. Some view grades as carrots and sticks—which indeed they always are, whether or not so intended—without any regard to the assessing and sorting functions. Other teachers, probably a minority in law but a larger percentage in other fields, see grades as actual measures of achievement on criteria that are absolute, not referenced against other students' performance. Such teachers adopt standards for each grade in the grading scale and then assign grades on the basis of whether students have exceeded the announced standards. In theory, their grades could be the same for all students in a class—all high, all low, or all in the middle. But even these teachers understand that their grades will often be read comparatively. And even if no one reads an individual teacher's grades comparatively, those grades will in all likelihood be combined into a GPA which will be used to make comparisons between students.[8]

7. It might be the case that standardization of grading would have the beneficial effect of making the process seem less arbitrary to students, while at the same time making it clearer that they are being graded in comparison to their peers. Such increased clarity might make it easier for a student to accept the disconcerting facts that he received a high grade on an exam that he did not feel good about and received a low grade on an exam that he thought he had nailed.

8. If (1) a teacher does not want to allow the possibility that her grades will be read comparatively to influence her criterion-referenced or incentive-based grading, (2) she does not combine subscores to get a total grade, and (3) her grades are not included in any summary statistic calculated by her school, she might find much of this article irrelevant to her grading.

2. Few readers know the grading practices of the individual teachers assigning the grades.

Employers and other readers of law school transcripts do not have the information or the time they would need to figure out what grades mean for individual teachers. A few hiring partners at local law firms might know that Softy gives high grades and Curmudgeon gives low grades, but most will not be able to make such individualized adjustments when they read transcripts or résumés. Many will be able to do no better than assume that the A- from Curmudgeon reflects a lower performance than the A from Softy, or that a B means the same thing to both.

3. I assume that teachers do indeed vary in their grading practices.

Even when they seek shared goals, even when they are partially constrained by rules and customs, some teachers award lower grades, some higher. Some spread the grades widely, some bunch them together on the grading scale. There is some evidence for this assumption that teacher-to-teacher variation has not been squeezed out of the grading process. Paul T. Wangerin finds that "dramatic differences in the definitions of letter grades exist within a single part of the university and even within different sections of the very same course."[9]

4. The impossibility of perfection in grading does not obviate error reduction.

I assume that grading is never perfect. If it were, many points of this article would be moot. But our assessment instruments will never be precisely able to measure student ability, or learning, or anything else relevant to performance as a lawyer. Subjectivity in grading will never be eliminated. Exam coverage may not be what we want it to be. Neither essay questions nor true/false exams give us an undistorted picture of the real abilities of our students to perform as legal professionals. No test is perfectly reliable or valid.

Even though there are large problems that can never be solved, we shall continue to grade students, giving them marks that they and others will read as meaning something about their accomplishments or abilities. I also assume, on the other hand, that grading is not so defective that improvement is pointless. We should take due care to prevent the errors we can eliminate and minimize the cost of the errors we cannot eliminate. The fact that our exams are not perfect measurement instruments is no excuse for saying that one student performed better than another when the opposite is true.

5. Schools will continue to sum incommensurable grades.

Because different assessment instruments measure different aptitudes and learning, a B in the four-hour Property course does not mean the same thing as a B in the four-hour course on Criminal Law. Yet, despite their differing meanings, the two grades have the same effect when schools calculate GPAs

9. Calculating Rank-in-Class Numbers: The Impact of Grading Differences Among Law School Teachers, 51 J. Legal Educ. 98, 112 (2001). See also Young, *supra* note 2.

and class ranks.[10] Even though different dimensions of performance cannot truly be aggregated into a single number, we are not about to stop doing it. And although GPAs have serious limitations, the faults are not so large that they cannot be made still worse. Some approaches to combining grades will exacerbate the problems inherent in summary statistics.

In addition to these assumptions, I will note that principles for constructing assessment instruments are beyond the scope of this article. I will not even begin to discuss the relative merits of oral examinations, drafting exercises, research papers, and essay and multiple-choice tests. Nor will I discuss the relative merits of points-based or gestalt grading. The discussion here is wholly about how we turn numbers, be they initial scores or grades, into other numbers, be they final grades, GPAs, or class ranks.

III. Principles of Comparative Grading

A. Equalize Means Across Courses

I start with the easiest point: the average grade should be approximately the same for all courses.[11] The reasons for this are obvious. Law schools commonly sort first-year students randomly into sections. It is plainly unfair for some students to suffer lower grades because of a throw of the dice, and especially so because first-year grades are the most important. That should be reason enough for teachers of first-year courses to follow similar grading practices.

After the first year, at most law schools, students choose their own courses. Allowing variation in the average grade in elective courses creates bad incentives for students. In order to improve their GPAs students might choose a course or section because it has higher grades instead of choosing according to important educational criteria, such as whether it meets too early in the morning. The incentive problem affects teachers as well as students. Allowing some teachers to reward their students more generously puts pressure on other teachers to raise their grades, perhaps resulting in spiraling grade inflation.

10. I have written elsewhere about the incommensurability problem inherent in creating grade point averages and class ranks from scores in different courses. Stake, *supra* note 3.

11. For good arguments in favor of equalizing means and empirical evidence that teachers in one law school do not give grades with the same means, see Wangerin, *supra* note 9, at 103, 109. There are arguments that medians should be used instead of means, but medians can be easier for a teacher to manipulate in undesirable ways. One thing a teacher can do is shift the grades of some students above the median to higher grades without shifting the median. This is hard if the school also controls the standard deviation, but it can be accomplished by shifting other grades downward. Another undesirable thing a teacher might do in response to a forced median is just change the grades of a few students near the median, rather than shifting the whole grading scale to meet the mandate. It is harder to shift the mean by changing the grades of only a few students because the amount the mean changes is the total shift divided by the number of students. If grading intervals are large, this is especially unfair to students whose grades were changed downward by the teacher just to make his grades meet the mandatory median. Notwithstanding these points, there are tradeoffs, and a school could rationally choose to mandate that all classes have the same median grade instead of the mean. Mandating the median and the mean along with the standard deviation amounts to mandating the skewness of the distribution of grades.

Variation in the average grade also results in unfairness and inefficiency when grades are combined. The students taking courses in which the teacher awards lower grades are less likely to qualify for honors, more likely to end up with a low class rank, and more likely to flunk out of law school entirely. In short, "the reliability and validity of GPA as a criterion of academic success are attenuated because the GPA is not comparable for students who take courses with severe grading standards and students who take courses with lenient standards."[12]

This equal-means principle should not be adopted unthinkingly or too rigidly. There is an argument that grades in elective courses should be higher than those in mandatory courses. Students in electives have chosen those courses and might be more motivated to study hard or might be particularly good at doing that sort of study. Electives are usually taken later in schooling, when students have matured. Especially when all students take the same mandatory courses, there is little harm in having one average for those courses and another average grade for elective courses. The grade average for courses could even vary according to the percentage of students taking the course as an elective.[13]

A further deviation may be warranted if there are a substantial number of pass/fail students in a course. Such students might not put forth even close to the same effort as students at risk of receiving a C or D, and their scores will probably be much lower than the scores of most other students in the class. If the letter grades given them, before being changed to pass or fail, are included in the average for the course, the instructor is effectively getting a free pass to give the rest of his students much higher grades than usual. On the other hand, there is also a potential fairness problem if those students' grades are excluded. They might have been the lowest students in the class even if they had taken the course for a grade. Eliminating their grades could force the teacher to give grades that are unfairly low to the regular students in the class. The best way out of this dilemma is to set the average grade for the regular students according to their GPAs in previous course work. The performance of the pass/fail students can then be judged according to the curve thus set by the regular students.

I have been asked to address the argument that we should allow teachers to give higher or lower grades because some teachers teach better. Certainly teachers are not equal in their effectiveness. We could allow the more effective teachers to give higher grades as a reward for teaching better. This would add to the incentives for good teaching, along with salary and other rewards. Students of better teachers learn more and should perhaps, as a matter of fairness, get better grades. It would also serve the goal of efficient communication to allow the better teachers to tell transcript readers that their students learned more than other students.

12. Stricker et al., *supra* note 2, at 178.

13. At some schools there are no courses that are elective for some students but required for others. At Indiana—Bloomington, however, we have courses required for J.D. students but elective for students pursuing other degrees; classes include students from both groups.

One obvious problem with this rationale for disparate average grades is that it is very hard to identify the teachers who teach better. Self-reporting will not do, as most teachers report that they are better than average,[14] and if there is some positive payoff to being a great teacher, most of us will report that we qualify. Unless a teacher can produce decent evidence of better learning on the part of her students, we should hesitate to accept such a contention as sufficient to warrant higher grading. And even if a teacher can produce evidence of better learning, it does not follow ineluctably that his students should receive higher grades. Law employers often want to know not how much the student has learned in law school but rather how capable the student is of learning, compared to the other students in her school. The student who learned less in the poorer teacher's course may be just as capable of learning as the student who learned more in the better teacher's course, contrary to the implication of the inferior grade. Thus, allowing the grade averages to vary according to the quality of the teaching might send the wrong signals to employers. Moreover, when a teacher changes his grading practice without warning and when the administration randomly assigns students to courses, the students do not have an opportunity to make informed choices. In such cases it would be unfair to the students to allow poor teaching to result in low grades in addition to weaker training than that of their luckier schoolmates.[15] In sum, although there is a logical case for allowing better teachers to give better grades, such a deviation from the equal-means principle would be unworkable, inefficient, and perhaps unfair.

Another argument for disparate averages is that the students are better in some classes. Few would object to giving better students better grades, but again there is a problem of proof. We all understand that what look like high scores on an exam do not necessarily mean that the students learned more than average. The test itself determines whether the students look like they learned a lot or only a little. In first-year or other mandatory courses, one teacher's randomly assigned students are not likely to be consistently better than another teacher's students. Substantial year-to-year variation in students is possible,[16] but it is also possible, indeed likely, for a teacher's test or grading

14. Patricia K. Cross found that 86 percent of a surveyed group of teachers considered themselves to be better than average. See Not Can, But Will College Teaching Be Improved? 17 New Directions for Higher Educ., Spring 1977, at 1, 5–6. It is possible that law teachers have better awareness of their own comparative teaching abilities than those surveyed, but I see no reason to believe it. It is also possible that student evaluations have pierced our self-deception since Cross did her study. There is also some small chance that Cross's results do not reflect self-deception or lack of information. If the bad teachers are much worse than most and most teachers are just a little below the best, a teacher at the median is well above the average.

15. The students of better teachers were lucky enough to be assigned to those teachers. Should the luck of the draw be equalized by giving them *lower* grades?

16. Substantial year-to-year variation is likely only in very small classes, or when admissions standards have changed. In the latter case, it could be appropriate for a school to change its mean grade for those students. Changing the target mean according to the aptitude of the matriculants would facilitate comparisons of students graduating in different years. If, however, the mean is increased, the problems associated with grade inflation, discussed below, might ensue.

17. This approach was adopted in the law school context at least as far back as 1971. See Richard

to vary from year to year. The teacher at my alma mater who tried to assign a huge number of D's and F's probably thought the students were much worse than those in previous years. It is much more likely, however, that his teaching, testing, or grading mood had changed. Small, ambivalent deviations from the norm by a teacher might be justifiable, but it is highly unlikely that one teacher will be randomly assigned better students over a long period of time. Even supposing, contrary to the odds, that one teacher's randomly assigned students are above average from year to year, allowing that teacher to grant consistently higher grades will raise a perception of unfairness and generate ill will among students. Since in law we care about the appearance of impropriety as well as its actual existence, we ought to avoid grading that looks unfair.

In the second and third years better students may gravitate to some courses, and lesser students to others. If that happens, forcing teachers to a single mean could easily increase miscommunication rather than decreasing it. For that reason, an exception to the principle of equalizing means might fairly be made for courses having students with different expected performances. If the students registering for one class have an average GPA of B+ and the students registering for another class have an average GPA of B-, the grades given by the two instructors could account for those differences.[17] Indeed, doing so could reduce fairness and incentive problems that can arise if some courses tend to attract students with high grades.[18] Again, where I went to law school, the course in federal jurisdiction was taken primarily by students who had high grades. The fact that the class would be full of good-grade-getters discouraged students from taking the course because they feared ending up at the bottom of the class. If the average grade for the course had been adjusted to account for the high GPAs of the students registering for it, any student could have taken the course without fear of harming his GPA. Although it might appear to students to be unfair that some courses receive higher grades, that appearance of unfairness could be prevented by explaining to students the fairness and incentive rationales for the policy.

This discussion would not be complete without an explicit exception for small classes.[19] It is always possible that forcing a teacher's grades to a prescribed mean can have the effect of increasing miscommunication. It is a matter of balancing risks. Is it more likely that the instruction or evaluation is abnormal or that the *class* is abnormal? The chances that some teacher-related component of the evaluation is abnormal are probably about the same for large and small classes, even if a teacher devotes less time to constructing the exam for the small class. By contrast, the chances that the class is actually abnormal increase as class size diminishes. Hence, the relative risk of error from forcing a mean upon the graders increases as classes shrink. At some

A. Epstein, Grade Normalization, 44 S. Cal. L. Rev. 707, 709 (1971).

18. *Id.* at 710 (a student's expected grade in a course should be the same as the median of his previous grades regardless of which course he takes).

19. Thirty to thirty-five observations is a typical rule of thumb used in statistics to distinguish between large and small samples. But the number of degrees of freedom and the nature of the sample are important too.

20. A 1996 survey done by Downs and Levit found only two law schools that constrain a teacher's

point before the class size diminishes to a single student, the costs of forcing a prescribed spread outweigh the benefits. Because small numbers of persons behave less predictably, a school's grading constraints might properly allow greater variation in the average grade in small classes. Over time, however, any given teacher's mean in such classes ought to fluctuate around the school average and not be biased in one direction or the other.

It should go without saying that there is no harm in dividing courses into groups and applying a different mean to each group. A faculty may wish to take some courses off the curve entirely, allowing all A's in writing courses for example. Care should be taken, however, to make sure that grading is consistent within each group of courses and that all students take the same number of hours in any group of courses. Moreover, if the means in upper-level courses are pegged to previous performances of the students, such means should probably be calculated without including the courses graded on any alternative scale to avoid creating an incentive for students to take extra third-year courses, or to take them out of order.

B. Equalize Standard Deviations

1. Equalize standard deviations across courses.[20]

Teachers should spread their grades to the same degree in all classes. One measure of variability in grades, as in any other set of numbers, is the range.[21] This is not a very useful measure, however, because it is based entirely on the top and bottom scores and ignores all the scores in between. A better measure of the variation is the standard deviation, which is essentially a measure of how far the grades fall away from the mean.[22] Since that is generally the best measure of variation if the means have been equalized,[23] to say that each teacher's grades should have the same degree of spread is to say that they should have the same standard deviation.

standard deviation. Robert C. Downs & Nancy Levit, If It Can't Be Lake Woebegone . . . A Nationwide Survey of Law School Grading and Grading Normalization Practices, 65 UMKC L. Rev. 819, 837 (1997).

21. "The range of a set of data is the largest value minus the smallest." It ignores the dispersion of the values between the two extremes. John E. Freund & Gary A. Simon, Modern Elementary Statistics, 8th ed., 73–74 (1992).

22. The population standard deviation, Σ, is the square root of the quotient of the sum of the squares of the differences of the scores from the mean score divided by the number of scores. This is sometimes also called the root-mean-square deviation. *Id.* at 75. The sample standard deviation, s, is calculated in the same fashion except that the divisor is n-1 instead of n. The sample standard deviation is used when estimating a population from one or more samples. The two statistics get closer as the size of the class increases. Although the two are not much different, it is interesting to think about which is theoretically better to use in the context of grading a class. Since we are not trying to estimate the actual standard deviation from a sample, but rather trying to equalize the spread of grades across different populations, the population standard deviation would seem at first blush more appropriate.

23. "The standard deviation is by far the most generally useful measure of variation." *Id.* at 74. Another statistic that could be used is the mean deviation, which is the average of the absolute values of the differences of the scores from the mean. The problem with this intuitive statistic is that the absolute values in its calculation lead to serious theoretical difficulties in problems of inference. *Id.* at 75.

To see the unfairness that can result when two teachers give grades with different standard deviations, consider the following example. Bert and Ernie are among thirty students, all of whom take two courses, Property and Torts. Ernie gets a 70 in Property and an 80 in Torts. Bert gets just the opposite, an 80 in Property and a 70 in Torts. Points are equally hard to earn in the two courses, and the number of exams is the same. The other students are the same students in both classes, and they perform exactly the same in both classes, so that the means and standard deviations on the exams are the same; 70 is the mean score and 80 is one standard deviation above the mean. Conforming to the first principle above, both teachers assign B's to students who achieve the mean score of 70. But the two teachers do not equalize the spread of their grades. The Torts teacher increases the grade by one notch (B to B+, etc.) for each one-half standard deviation above the mean. Ernie scores his good grade in that class, so he gets an A-. The Property teacher spreads grades less broadly, increasing a grade by one notch for each full standard deviation. Bert's good grade is in that class, so he gets a B+. When the two grades are averaged for the GPA, Bert's GPA will be 3.15, and Ernie's will be 3.30.[24] The same total performance results in different grades and quite different GPAs. Although this example involves a numerical scale, the same principles apply if the grading is by the gestalt or any other method. A similar example could be constructed to show that meaningfully different performances can result in the same grade.

In addition to this horizontal inequity of like performances receiving unlike grades, notice that the teacher who gives a wider spread has a greater impact on the GPAs of the students. A teacher giving grades with a large standard deviation puts more students into a position to receive special honors (or dishonors) than a teacher who does not spread scores so widely. Conversely, a teacher who grades with a smaller variance gives fewer students a chance at honors. Unless a faculty wants to allow some self-selected teachers to award more of the accolades and push more students to the bottom of the class, the faculty should set a standard deviation for all ordinary classes. Of course faculties willing to so constrain their grading must find a way to make it easy for teachers to know the standard deviation of the grades they give, but software is available for that purpose.

Grades with unequal variation also create bad incentives for students choosing their electives. A student who generally does well will choose courses in which the spread is wide, if she knows enough to do so, because there are a lot more A's to be had than in courses in which most grades are bunched around the mean.[25] On the other hand, a student who generally struggles has an incentive to choose courses with little variation because those courses make it easier to get an average grade, or close to it. Another incentive problem,

24. A real student's GPA is always constructed from more than two grades; I chose this example to keep the math simple.

25. This assumes that the courses have the same mean grade.

suggested above, is that students taking courses with different grade variances would focus their study efforts on the classes with the larger variances.[26]

One objection to equalizing standard deviations might be that, as a practical matter, it is just not important enough to be worth the costs of doing so. On the one hand, the costs borne by teachers are not high because inexpensive grading software makes the task a snap. But what about the other hand, the costs of unconstrained variance? To determine whether forcing teachers to give grades with equal means and standard deviations would make any actual difference to students, Paul Wangerin studied one law school's grades over eight years. He found that such a requirement would have made a difference of up to 40 places in the class rank of some students. In addition, three persons who were not ranked number one in their classes would have been the top students if the teachers had equalized their grading.[27] Those inequities seem large in light of the ease with which standard deviations can be equalized.

The arguments above support a rule that requires all teachers to give grades with the same standard deviation, and it is certainly possible for teachers to come close to a prescribed standard deviation in all classes of reasonable size, whatever the spread of actual performances. But there is a problem with mapping narrowly bunched performances onto widely spread grades. If the students' performances appear to be about the same to the teacher, she will have little confidence that she has ranked the students in the proper order. If the ranking is wrong to start, magnifying the differences to reach a mandatory standard deviation will increase the unfairness and inefficiency of the grading.

Observed student performances might be closely bunched for a number of reasons. It is possible that the actual student performances in a certain course do not spread as widely as in the usual class. We can never know this is true, but it will be true in some cases. At this point, we must again acknowledge that forcing teachers to conform to a prescribed statistic can have the effect of increasing miscommunication. Because small groups are less predictable, a school's grading constraints might properly allow greater variation in the standard deviation for small classes.[28] As with means, however, any given teacher's standard deviation in small classes ought to fluctuate over time around the school's average standard deviation and not be biased in one direction.

26. One might ask how students will find out that there is more variation in the grades of some teachers. That is easy to do if grades are posted. Students can get a rough measure of the variation by simply counting the A's. Even when grades are not posted, students can compare notes and find out that there were a lot more A's in some courses than in others. I knew that my tax teacher awarded A's to only three students in our class of about 250. Some students in the following classes also knew that effort counted for little in that course. Those students who were not aware of the pointlessness of effort in that class may have been arbitrarily penalized in their grades in other courses on which they spent relatively little time.

27. Wangerin, *supra* note 9, at 114.

28. Once again, there is nothing wrong with dividing courses into groups and mandating different standard deviations for each group as long as all students take the same number of credit hours of courses from each group.

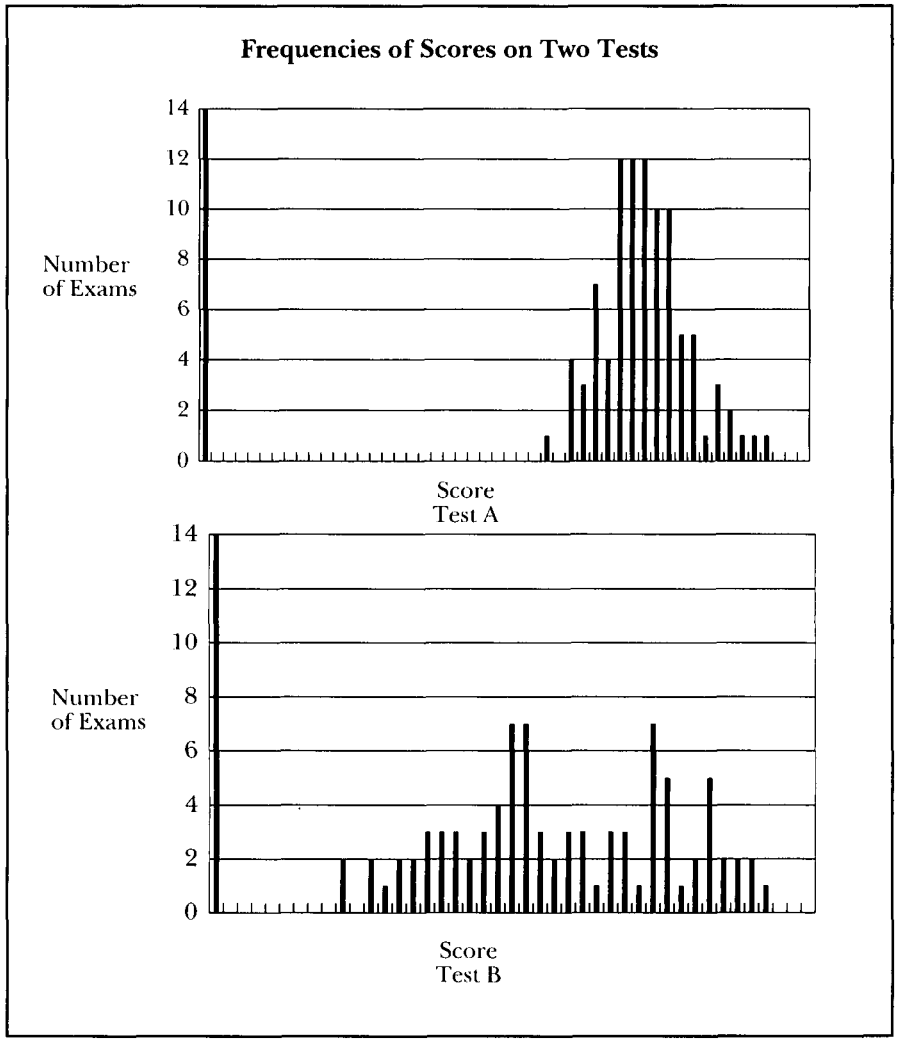
Student performances can also appear to be tightly bunched if the evaluation instrument is poorly designed. If all multiple-choice questions are either very easy or very hard, students will tend to have the same scores because all students can correctly answer all the easy questions but no students can correctly answer the hard ones. Or, on an single-essay exam, the question might not elicit much of a range of responses from the students, all seeing the obvious points and few perceiving the subtleties. These situations present a real problem. Standard deviations should be equalized but, on the other hand, with no differences on the exam the teacher has no reliable basis for making distinctions. What should be done when a teacher's evaluation instrument yields little in the way of meaningful information? Should he be allowed to deviate below the prescribed standard deviation, effectively devaluing his grades and the importance of his course, perhaps if he is willing to write a letter to the administration explaining the deviation from the norm? If students were randomly assigned to his course, or if they elected to take the course without a warning that it would be graded differently, his written explanation will do little to compensate for the unfairness of their reduced opportunities for high grades. Only if the students choose the course after notice that it will have a lower variance is it fair to allow the teacher to give them less chance at receiving grades that would help them to achieve honors. When a teacher administers a poor test, he creates problems that are not easily solved.

A more defensible reason for allowing teachers to deviate from a prescribed standard deviation is that some courses for some reason appeal only to a narrow range of students—the top, bottom, or middle students. As I suggested with means, an exception might be made when the teacher knows that the GPAs of the students enrolled have a smaller standard deviation than the GPAs in similar courses.

Some teachers might feel that their grades deserve more weight because their tests spread students more widely than the average test. While it is certainly true that some tests are more reliable instruments of assessment than others, this argument conflates the scoring of exams with the scaling of the scores. Scoring is, for good graders, not arbitrary. Good exams garner high scores and lesser papers draw lower scores. But scaling *is* arbitrary. With the exception of teachers, if any, who actually apply absolute standards,[29] all grading involves mapping raw scores (or even impressions, for those who do not use numbers) onto a grading scale. Suppose that one teacher's test yields three scores, 33, 30, and 27, while another teacher's test yields 18, 15, and 12. The latter test did spread scores more broadly in a percentage sense, but the scores on both tests have to be scaled, and the standard deviation of the

29. One example of an absolute standard is the old-fashioned standard of "good," which was sometimes equated with a B, which was equal to a 3.0. This assumes, of course, that we know what "good" means. Of course one could argue that even the standard of "good" is relative rather than absolute. But such an argument only cuts in favor of the point made here, namely that grading is often, if not always, comparative, and therefore whether we map test scores onto higher or lower, or broader or narrower, points on our grading scale is in many ways arbitrary.

resulting scaled scores on most grading scales will be lower than that of either test.[30] What is the right amount of reduction in the standard deviation? There is no right answer. It is completely arbitrary. Moreover, there is no reason to believe that the latter test is in some way a more reliable instrument. It is possible, for example, that the former test could be turned into the latter test by subtracting the right fifteen questions. While those fifteen questions might be doing nothing to add to the power of the test to discriminate between students, those questions are not harming the discrimination either and they do gather evidence about the knowledge of the students. Hence, the greater percentage spread in the actual scores provides no basis for concluding that one test deserves more weight.



30. The sample standard deviation, s, of the scores on the two tests is 3. (The population standard deviation, Σ, is 2.45.) The standard deviation of grades in most law courses at Indiana is less than .5. To fit our scale, the test scores would have to be squeezed down by about a factor of 6 (or 5).

To probe the claims just made, I selected fifty-one problems from my 2002 Property exam and called those problems Test A. I selected a different fifty-one problems from the same exam and called it Test B. Then I rescored all of the students on the two exams. The average grade on Test A was 35.1 and on Test B was 31.6. The standard deviation on Test A was 3.5, but on Test B was 8.8! The figure shows a graphic depiction of the different distributions of scores on the tests.[31]

A teacher might look at the much larger standard deviation on Test B and say that the students who took Test B should have their grades spread more widely than the students who took Test A. But these are the same students tested on nearly the same material at the same time![32] It is highly unlikely that the variance in their aptitude or effort was different and, whatever the difference in spread due to the difference in the substance of the questions, it stretches credulity to think that the students deserve a substantially different spread of grades on the two tests.

The general point here is that what appears to be a wide spread of scores on an exam provides no basis for spreading grades more widely than normal. It is entirely possible that the wide spread is purely an artifact of the test or scoring procedure and not a real difference in the underlying levels of aptitude or learning. Unless there is external evidence comparing the students in one class to the students in another class, the spread should be the same. The results of a test given only to one class, standing alone, provide absolutely no basis for spreading grades more, or less, broadly than average.

Suppose Professor Puffedup is particularly proud of his course and argues that his grades should have more impact because his course is harder than others and imparts more learning than other courses of the same credit. We, Puffedup's colleagues, cannot deny that some four-hour course has the greatest content, however defined, and Puffedup insists that his course is that course. Puffedup's argument loses some force, however, when we recognize that he is essentially rehashing a point that the faculty has already decided, a point that was settled when the faculty gave his course (and others) four hours of credit toward graduation and four hours of weight in the GPA. Given that truce, the issue is whether the school should allow Professor Puffedup to unilaterally say, "My course counts for more." My guess is that most faculties would deny him that authority. Moreover, even if we agree with Professor Puffedup that his course is worth more and therefore do not mind that he gives his grades more clout, we should still be bothered by the fact that his deviation makes it possible for equal total performances by two students to result in unequal class ranks.

Another claim that I have heard a teacher make is that he writes and grades exams more reliably than the average teacher, and therefore it is not a

31. The scores along the horizontal axis increase by one point for each hash mark. The bar on the left of each figure represents no data and should be ignored. It was added to equalize the scale of the two figures.

32. The material for both tests was taken from all parts of the original exam. But since the correlation coefficient for the two tests was only .48, it is clear that the two tests did cover different material or skills.

problem that his grades count more heavily when grades are summed into GPAs. It is true that some teachers' grades are probably more reliable than others', but which teachers? Once again, we confront a problem of proof. Unless a teacher can convince his colleagues that he is better at assessing performance, he should be constrained by the same limit on grade variation as constrains his colleagues.[33] He might argue that neither he nor any other teacher need be constrained; others are free to increase their spread if they want their grades to have weight equal to his. The defect in his argument is that responsible graders might not have the freedom he feels. Teachers who believe that grades should be proportional (discussed below) and believe that they should not give failing grades to students with passing performance may not be able to expand their grades enough to equal the variation of the teacher who does not feel constrained by those two principles.

A more egalitarian variant of the argument above proceeds from the premise that teachers vary in their grading acuity from year to year. One year's exam generates responses that allow the teacher to feel very confident in her grades. The next year's exam in that same course generates like responses from all students in the class, leaving her with no confidence that she has reliably detected the variation that actually existed in the class. To take account of this varying reliability, we could allow teachers to vary their standard deviation from the target in any given year as long as they hit the target, on the average, in a given period of years. Grades in which teachers had more confidence would get more weight in the average, while grades in which teachers had less confidence would acquire less weight. This should make the GPA a better predictor or measure of overall student performance.

This approach would raise serious issues, beyond the obvious but surmountable administrative difficulties. As I pointed out above, diminishing the weight placed on the grades in one course could be viewed as unfair to students who did well in that course and unfair to the students who did poorly in another course. The deeper unfairness, that of treating equals differently, does not arise here because the premise of varying exam quality denies the equality of performances that was presumed above. But unless the course was an elective and there was notice that it would have a lower variance, it might still seem unfair for one course to have an unusually large or small impact on students' grade point averages. Since a system that allowed teachers to vary around a target standard deviation would prevent teachers from grabbing[34] more weight for their classes and since it could be more accurate, we face something of a tradeoff between accuracy and fairness. The unfairness here might be eliminated by telling the students in advance that courses will be weighted according to both the course credit hours and the quality of the

33. An interesting issue is presented when a teacher can produce good evidence that her exam is more reliable than average. In such cases should the school give that teacher's grades more weight? Doing so would increase the accuracy of student GPAs, but it would create incentive problems if it were known to the students and fairness problems if it were not known.

34. Such a system would also prevent a teacher from unintentionally giving her course more weight in the total.

exams as determined ex post by the teachers.[35] A student in the ex ante position could rationally prefer the more accurate but less predictable system, and in that sense it could qualify as fair. It also seems likely that employers would prefer that system. But the students would probably feel something like the caged rat that never knows when he is going to suffer his next electric shock. My guess is that reducing the predictability would generate more anxiety than the increase in accuracy was worth.

So the standard deviation should be equalized across classes. But we can justify limited exceptions to allow leeway for small classes, to account for differences in prior performances of the students, and to minimize the impact of poor evaluation instruments if students are properly warned of that possibility in advance.

2. Be careful when forcing grades to a curve.

The argument made here for standardization should not be read as an argument for normalization. Some schools attempt to solve the problem of unequal spreads across courses by mandating that certain percentages of the class fall into each of the available grade intervals. If the percentage ranges for each interval are very narrow, the forced curves will have the salutary effect of keeping all teachers to the same mean and standard deviation. The forced curve has the added benefit of avoiding differences in skewness and other statistics that are used to describe distributions.[36] Nevertheless, there are a number of potential problems with specifying the precise percentages of grades to be awarded in each of the grading intervals.

The first problem is that the mandated curve is often based on a normal distribution, and the students in a school might not actually be normally distributed about the mean in their combined aptitude and effort. In fact there are good reasons to believe they are not normally distributed. Assuming that legal ability is normally distributed in the population, it is likely that for the most part only the top portion of the population is admitted to law school; the law schools may make up a collective Lake Wobegon in the sense that all students are above average. So the distribution of students might look something like the upper half of a bell curve, but in any case it seems doubtful that it is normal. If all of those students were to be divided equally, each school's curve would be expected to be nonnormal, perhaps something like the right half of a bell curve.

35. If all teachers go high or all teachers go low on the standard deviation, allowing that freedom has not really harmed the process. If some go high and some go low one year, the opposite will be true another year, and in both years the more accurate exams will get more weight. But that will make comparisons across classes more difficult.

36. Roughly speaking, a distribution is skew when the tail on one end is longer than the tail on the other end or, more accurately, when it is not symmetrical about the mean. The Pearsonian coefficient of skewness is calculated by subtracting the median from the mean, multiplying by 3, and dividing by the standard deviation. There are other statistics designed to describe distributions, and they could also be equalized across classes (and would be equalized with a forced curve), but trying to equalize them might not be worth the effort, although I know one teacher who has done so.

Second, applications and admissions differ across schools. Starting with a group of mostly above-average students, some schools' admissions processes cut off the lower end of the distribution fairly abruptly, ruling out the poorly credentialed students and admitting many students just above the cutoff. Of those students that barely made it, many will choose to attend because they will have few options that are better; they will not qualify for admission to more selective schools. At the top end of the admitted students, the percentage that matriculate will be low because those students have many options. The top of the distribution is thinned out in the fierce competition with other schools in a way that the bottom of the distribution is not, exacerbating the skewness of the underlying population.

In short, at many schools a disproportional number of the students admitted will be at the lower end of the school's range, and among those admitted a disproportional number at the lower end will accept admission.[37] Of course admissions policies vary across schools, and the distributions of students will vary accordingly. The point here is not to try to predict what the distribution of abilities will look like at all schools, but rather to point out that there are good reasons to expect that the students will not be normally distributed in their combination of aptitude and effort. A group of nonnormal scores can be mapped onto a normal curve only with a nonlinear transformation,[38] the problems of which are discussed below. For those reasons, it could well increase miscommunication for a school to require all courses to have the same number of A's as C's, or have grades that are symmetric about the median or mean, or have tails as long as those in the normal curve, or otherwise conform to any single predetermined distribution.

A school could account for the preceding deviations from normality by determining a local curve based on its own students. It could "grade" its students on the basis of their intake statistics (usually the undergraduate GPA and LSAT score) and use the resulting distribution as the template or curve to be applied in each course.[39] But there are still other reasons to worry about whether the students in any particular course will be a representative sample of the whole school. One reason we might not see a school-typical distribution is that some students might not have prepared at all, making their performance much worse than those ranked just above them. Another reason the performances might fit the school's curve is, as mentioned above, that a particular course might attract a special group of students, either above or below the mean or perhaps less widely distributed than the usual class in that school.

37. Thus, although law students as a group are above the average of the entire national population, the majority of the students at any one school may be below average for that school.

38. A teacher at one school with a forced curve reports that it sometimes prevents him from maintaining proportionality in his grading. By contrast, means and standard deviations can always be equalized with a linear transformation.

39. One problem with this approach is that the UGPA and LSAT score, even combined, do not predict grades reliably enough.

These last problems could be partially mitigated in upper-level courses by tuning the class distribution to the actual performance of the students in previous course work, a curve based on the actual GPAs of the students in the class. But this is not a solution either. If the mandate is highly specific, it could force teachers to give students inappropriate grades. It is likely that a curve requiring "2 to 3 percent grades of D" will force a teacher to give more or fewer D grades than the class actually deserves.[40] There will be test-to-test variations in the performances of the students that cannot be accommodated by any precise grading curve. Of course the same concern could also be used to argue against a mandatory mean or standard deviation, but the point does not apply with equal force in that larger context. The possibility that the relative student performances do not conform to predictions is greater when the numbers are smaller.[41] For a class of two students, any mandate, whether mean, standard deviation, or curve, is dangerous. For a class of 500, all three can be mandated without a problem because there will be a large number of students in each interval.[42] For a class of thirty students, however, forcing the scores onto a curve will be troublesome because some of the intervals will contain very few students. For classes of the size that are common in law schools and for grading scales with more than a few intervals, forcing a specific percentage of students into each interval is riskier than mandating a standard deviation because each of the intervals involves only a portion of the class, whereas the standard deviation relates to the whole class. With the whole class, small variations will average out in a way that is not likely when an interval includes only a few students.

Some schools have accounted for this problem by loosening the mandate, allowing teachers some latitude in the percentages of grades falling into each interval. While that solves the small-number problem, it recreates the original problem of variation in grading across classes. When all the small deviations in each grading interval are added up, the results may be substantial variation in the means or standard deviations from teacher to teacher.[43] The flexibility needed to account for irregularities in the actual distributions of performances allows too much variation in the mean or spread of the grades. By contrast, mandating a mean and standard deviation keeps variation in course importance to a minimum while reducing the specific inequities that arise when a curve is imposed on a distribution of performances that did not actually fit that curve. A teacher can reach a specified mean and maximum standard deviation even when the distribution of scores on her test is bimodal,

40. One of the reasons that this is a problem is explained in the section below on proportionality.

41. The small-numbers problem is a function of the number of the grading intervals and the number of students because the total number of students will be divided by the number of intervals. For reasons discussed below, intervals should be numerous. But the more numerous the intervals, the more a forced distribution is problematic.

42. Mike Alexeev pointed this out. I admit I am not accustomed to thinking about such large classes.

43. Some schools fix this problem by constraining the mean and standard deviation as well as the distribution. See Downs & Levit, *supra* note 20, at 838. But those schools may run into the small-numbers problem if they attempt to fetter the distribution too tightly.

single tailed, or otherwise out of step with the school or the previous performances of the students in that class.

Despite the points above, it would be rational to take the position that a school needs a mandatory distribution to counteract the effects of bad examinations. Suppose a bad exam has artificially created a bimodal distribution. If the exam preserved ordinality, although it lost cardinality, a forced curve can put the students back where they belong in relation to each other. If the forced curve is based on the predictors of performance for the actual students in the class, and some freedom is given to account for variation in small numbers, the curve might do more good than harm. But if the curve does allow leeway, it will not fix the mean and the standard deviation. Therefore, unless a curve is specified narrowly, it should not be seen as a complete substitute for a forced mean and standard deviation. If the class is large enough to apply a curve, it is more than large enough for fixing the mean and the variance.

3. Schools can use standardized grades to rank students.

Some faculties would object that forcing means and standard deviations upon teachers violates their academic freedom[44] or undermines the goals of criterion-referenced assessment. Other faculties will find other reasons not to force teachers to conform their grading to predetermined statistics. But even in such schools, all is not lost. As is suggested by Wangerin's study, the goals embedded in the first two principles above might be partially accomplished by a school's administration without forcing teachers to change any of their grading practices. Rather than calculating actual GPAs, the administration could calculate a summary performance score (SPS)[45] and a class rank in the same way that many careful teachers calculate grades from multiple components, that is to say, using Z-scores or T-scores. In other words, the administration could turn all teachers' grades into Z- or T-scores and use those for calculation of the SPS and rank.

What are Z-scores and T-scores? They are scores that have the same means and standard deviations and are often used to make it easier to compare or combine scores on different tests or parts of tests. A Z-score is created by subtracting the mean and then dividing by the standard deviation (Z-score = (observed score - average score) / standard deviation).[46] The mean of a set of Z-scores is 0, and the standard deviation is 1. Z-scores are scores for which the means and standard deviations have thus been equalized.

Because about one-half of the Z-scores are negative, some teachers prefer to use T-scores. T-scores can be derived from Z-scores by multiplying each score by 10 and then adding 50 (T-score = Z-score x 10 + 50). The resulting set of scores has a mean of 50 and standard deviation of 10. It makes no mathematical difference whether an administration combines grades using the Z-

44. See *id.* at 848–52.

45. To avoid confusion with the GPA, this SPS should be on an obviously different scale from the grades.

46. For additional discussion of Z-scores, see Wangerin, *supra* note 9, at 102.

or T-score method. Both have "standardized" the component parts of the overall performance score according to the standard deviation of the grades in the courses, and both give equal weight to all courses.

Once it has created Z-scores or T-scores out of the grades in the courses, the administration can multiply those scores by the number of credit hours in the courses to give each grade appropriate weight. After that, the products can be summed or averaged to yield a single overall performance score. The administration would thus treat each teacher's grades as one test in a series of exams being given the students. The final scores could then be used to determine class rank.

There are actually a couple of advantages to accomplishing this equalization administratively. An individual teacher of a small class graded with a coarse grading system may have a hard time conforming her grades to a prescribed mean or standard deviation. If her average is too low, for example, moving only a couple of students with the same score to a higher grade might make her average too high. The administrative approach allows her to give whatever grades she feels appropriate and does not require her to do any mathematical or other adjustments. The administration does the mathematical manipulation for her and, for purposes of determining class ranks, can in effect award students grades that do not exist on the grading scale in order to make her grades comparable to grades from other teachers. The other advantage for faculties uncomfortable with grading mandates is that teachers are not forced to grade according to a curve. They can assign grades on whatever basis they choose, letting the administration attempt to make the grades compatible when summing them up.

This approach will not work in all situations, however. To take a simple—albeit extreme—example, if a teacher gives only one grade, there is no fair way for the administration to give a set of scores that varies as required to meet the expected standard deviation. In less extreme cases where teachers have given mostly one grade with only a few grades higher or lower, it is perhaps improper for the administration to spread them more widely.[47]

Another way for the administration to reduce grading disparities is to do the scaling for the teachers. At my school, the administration offers to take raw scores from teachers and combine them using the weights for those scores that the teacher provides. Because the teacher retains final control over the mean and standard deviation, this approach is less intrusive, and less effective, than administrative rescaling of all the grades recorded by the teachers. Nonetheless, the process can help keep teachers from deviating too far from the practices of other teachers by suggesting a reasonable set of grades.

Assuming an administration does use Z-scores for combining grades, is there any reason for individual teachers to equalize means and deviations? A couple of reasons remain. One is that students might compare grades and be

47. As discussed above, an assessment that results in almost all scores being the same can raise problems that are not easily solved. When a school intends to include a course in GPAs and class ranks, the teacher should be encouraged to assess performance in ways that yield meaningful differences.

misled if the actual means and deviations have not been standardized across courses. Similarly, some employers might compare two students' grades (perhaps students from different years so that class rank is not helpful) and get the impression that one is a better student when the truth is the other way around. To cure this, the administration might rescale all scores and report those rescaled scores on the transcript. The faculty might object, however, that grades should be given by teachers, not the administration. Nonetheless, if the administration were to calculate the SPS and the rank by using Z-scores, the need for teachers to submit grades with uniform means and standard deviations would be substantially diminished.

4. Teachers should use standardized scores when combining components of a course grade.

Even if the administration uses standardized scores to combine grades from different courses, it is still important for teachers to follow the principle of equalizing standard deviations if they create composite grades from subscores. Most teachers grade on the basis of more than one question or set of questions.[48] Some teachers do not try to accord their scores different weights, they just add them up. Other teachers go to some trouble to give different weights to different parts. Many teachers do not give students a clear indication of the relative weight that will be placed on the various components that make up the final grade. The upside to that omission is that the teacher is not sending false signals no matter how he combines the scores. Another advantage is that the teacher learns more about what the students know without studying. The downside is that the teacher learns less about what the students could know if they had a chance to study.

Although I specify the weight for each set of questions on my exams, I do not take the position that all teachers should do so. If a teacher does make representations regarding the process of calculating final grades, she should live up to them. She should do so for simple reasons of honesty, of course. But that is not the only reason. Once again, the inefficiency and unfairness of miscommunication are implicated. Take an example from outside of law teaching. Professor Athlete teaches a clinical course on track and field. He announces that 95 percent of the final grade will be the student's time in a 100-meter dash. The other 5 percent will be his height in a pole vault. Then, when he grades, Athlete gives 5 percent of the weight to the dash and 95 percent to the pole vault. Sensible students will have spent most of their time practicing the dash. As a result, the pole vault scores will not be a very reliable indicator of the students' abilities to do the pole vault, much less their overall athletic abilities. While some students might never have improved, some might have been able to do much better if they had allocated practice time to pole vaulting. The scores on the 100-meter dash will be much more reliable, but they are given so little weight in the final grade that those grades end up being mostly a report of the unreliable information gained in the pole vaults.

48. Teachers who do not combine subscores into some form of total score can skip ahead to the next section, III C.

Thus the grades are not good indicators of track and field ability. Certainly, if a coach were trying to choose a good team on the basis of grades, she would prefer that Professor Athlete weight the scores as he had said he would.

Returning to the typical law school course, the more the teacher misleads a student about what will be on the exam, the less she will study the material tested and the less reliable the test will be as a measure of the student's ability to learn. The more the teacher misleads a student about the weight that will be given each portion of the exam, the greater the possibility that she will receive a low score because she spent too much time on one problem—not because she could not have answered the other problems well. Teachers should assign grades as promised.

Many teachers tell the students in advance how much weight each question will receive in the final accounting.[49] Sometimes they succeed. But other times they end up giving more or less weight than they and the students thought would be accorded that portion of their assessment. The key to getting weights right is paying attention to the degree of variation in the scores on each component of the total. Failure to account for variability of these scores can undermine the best intentions to weight carefully and provide accurate information to the students about the basis of assessment.

To see how such a failure could occur, take an example. Suppose the course grade is to be made from two subscores, each having equal weight. Each part of the exam has a total possible score of 100. On part A, the scores are normally distributed from 55 to 65. On part B, the scores are normally distributed from 20 to 40. Having taken statistics, the teacher remembers that he cannot simply add the two scores together because the students obviously performed differently on the two parts. To live up to his promise of equal weight, he multiplies the scores on part B by 2 so that the average on part B is equal to the average on part A. This is a huge mistake because the scores on part B now run from 40 to 80. When the final grades are given, most of the students in the class will receive the same grade they would have received if part B had been the only assessment instrument and part A had not been counted at all. If you doubt this, just imagine that the scores on part A had run from 59 to 61 with an average of 60. Will part A have as much impact as part B now? The effect of the teacher's method was to further devalue the scores that were already devalued by the process used for scoring the two parts. That is neither what the teacher wanted nor what the students expected.

49. In this discussion I do not mean to prejudge the question of what is promised. A teacher could say that 10 percent of the points on the exam will be on section A. If she gives 10 possible points for section A and 90 points for section B, she has fulfilled her promise, even if all students get all 10 points. On the other hand, if she says that 10 percent of the variance in the grades will come from the difference in performances on section A, she cannot fulfill her promise if all students get all 10 points. When a teacher says that "section A will be given 10 percent of the weight," it is not clear which she means, points or variance, but I will assume the latter meaning. I take no stand on what teachers should promise, but whichever is intended, the teacher ought to make the intention clear to students, or at least try to leave all students with the same understanding. I also take no position on what a teacher should do when she has written an exam that makes it impossible for her to live up to her promise, but I acknowledge that giving large weight to minute differences in performance is fraught with potential for unfairness and inefficiency.

Readers who have slogged through the preceding discussion of standardizing the spread across various courses will recognize that the solution is to equalize the variation in the scores in those parts. If the amount of variation can be made the same, the parts will have equal weight in the final summation. The teacher should linearly manipulate the scores so that the scores in each subpart have the same standard deviation. This is a simple matter of dividing each score by the standard deviation for the scores in that part.[50] After that division, the scores on each subpart will have a standard deviation equal to 1, and therefore will be equally spread out and consequently will have equal weight.[51] Of course, standardized scores indicate only a student's relative performance in the class. If there is some absolute level of performance that can be identified in advance as being important for some reason, that level must be attended to using the raw, unstandardized scores, or that level has to be standardized using the same linear transformations before the standardized scores are compared to it. By equalizing the variance in their subscores before combining them, teachers can produce grades in a way that conforms to the expectations created by their syllabi and announcements.

C. Use Smaller and More Numerous Grade Intervals

One of the common misconceptions about grading is that it is better to use a coarse grading scale with fewer and larger groupings because the grader can have more confidence in the grades. Clearly it is true that no one will get the wrong grade if one grade is given to all students. Likewise, teachers probably make few errors when they have to determine only whether the students pass or fail. Generally a teacher will make fewer errors if she uses fewer grades than she will if she tries to make finer distinctions. But that is only half of the story, and in my view the less important half. There is a tradeoff between, on one side, the frequency of errors and, on the other side, both the precision of information recorded for subsequent retrieval and the magnitude of the errors in that information.

To take a simple case, suppose that grades in one class are based on a single test, the class's scores range between 5 and 13, the standard deviation in the scores is 3, and the standard error for the test is 1.[52] That standard error means

50. I subtract the mean before dividing, thus turning the raw scores into Z-scores. One of my colleagues refuses to use Z-scores on the grounds that even this simple transformation diminishes transparency in grading, making it harder for students to understand their grades and for him to check the results for arithmetic errors. Transparency is valuable, but I think it would be better for schools to achieve it by explaining the transformation to the students. As for math errors, good computer software should not make any, but I admit that the Z-score method does make it somewhat harder to catch a computer's errors. I allow my students to see both the raw and Z-scores so that they can check the final results to see if they make sense.

51. Fortunately there is computer software available for making all of these calculations. All a teacher has to do is fill in the scores on the subparts, supply the weights for each subpart, and supply the institution's grading scale. The software will do the rest. If the class is not too small and the exam does a decent job of identifying differences between students, the resulting grades will be weighted appropriately, will have the desired mean and standard deviation, and will be proportional to the institution's grading scale.

52. Real exams usually have more points and larger standard errors, but the same principles apply. For a real example, see *infra*, page 608.

that the teacher can be somewhat confident that the test has not mismeasured each student's true score (real learning or ability) by more than 1 point. In other words, the student who scored 6 might more accurately be described with 5 or 7, but it is unlikely he deserved 4 or 8. If this teacher's school allows her to give grades from C- to A+ and she maps each of her students' scores of 5 through 13 onto those nine letter grades, she and the school cannot have much confidence that any of those grades is correct: a number of them will be wrong. The school can reduce the number of inappropriate grades by limiting the grading scale to three grades, A, B, and C, where 5–7 = C, 8–10 = B, and 11–13 = A. The students who scored 6, 9, or 12 probably get the right grade, and students who scored 7, 8, 10, or 11 are only half as likely to receive the wrong grade as under the finer grading scale. The coarser scale reduces the frequency of errors.

Two problems are created when a school reduces the frequency of errors by imposing a grading scale that is coarser than that used to score the tests or other measures of performance. First, information is lost in the process of rounding the scores to the nearest marks on a coarse grading scale. When the teacher herself looks back at the grade later, she will not know whether a student she gave a B scored 8, 9, or 10 on the exam. That imprecision in recording is added to the imprecision of the test itself, with the result that the teacher can be confident only that the B student had a true learning of somewhere between 7 through 11. If she had recorded the grade as a B-, she would know when she looks back later that the student's true score was probably 7, 8, or 9. Other readers of her grades are certainly in no better position than she is to interpret her grades. If a school requires teachers to map detailed information onto a coarser grading scale, the school forces teachers to throw away information that cannot later be reconstructed by any reader of the grade.[53]

This problem—that coarse grading scales conceal meaningful differences—can also occur when coarse grades are combined into an overall GPA if there is little variation in the student performance and little measurement error. If Joan's scores in three courses are 10 and John's three scores are 8, they will both have a B average even though Joan's performance total of 30 is much better than John's total of 24. Even if 10 is not significantly different from 8, 30 can be significantly different from 24, but the reader of the transcripts will never know of the meaningful difference in performance. If, however, the grading scale allows the teacher to modify each whole grade with a plus or minus, the differences in total performance will easily show through. Joan will have a B+ average and John a B-. So the first point in favor of a scale with finer gradations is that it allows teachers to record and communicate more information than can be contained in a coarser scale.

It is true that this problem diminishes as variation in the student performance increases relative to the width of the grade intervals and as the number of courses being averaged increases. Whether there is enough variation to

53. Of course the teacher can keep finer information if she wants to, but readers of transcripts usually do not have access to such information.

allow differences in students to show through the coarse grading scale becomes an empirical question. In the first year of law school, where grades matter most, each student's average is composed of only a few grades. In schools with few grade intervals, students with differing performance might have the same GPA, but perhaps that does not happen often enough to matter. I do know that a coarse grading scale led to fifteen number-one students in the class of 2002 at my daughter's high school. Like actual student variation, error in the measurement instruments may also make it easier to distinguish between students, but that possibility should give us no comfort as it raises the next—more serious—problem.

The second problem created by coarse grading is that errors are nastier. Continuing the example from two paragraphs above, some students who scored 7 had a true score of 8. They should have gotten a B instead of a C. By giving a C, the teacher has said that these students are well below average when she should have said they are average. Conversely, some students who scored 8 had a true score of 7. They should have gotten a C instead of a B. Under the coarse system the errors in grading are large, a full grade. Under the finer scale the errors are more frequent but much smaller; there is little chance that any student's grade is off by a full standard deviation.

An actual example can be constructed from the Property test I gave to ninety-four students in the spring of 2002. The standard error for the exam was about 5, and the scores ran from the 60s up into the 130s, with a mean of 96.4.[54] Suppose I divide the scores into three intervals: C's below 82, B's from 82 to 111, and A's above 111. Looking at the students' actual scores, I would have sixteen persons within one standard error of the A-B border and seventeen students within about one standard error of the B-C border, for a total of thirty-three students for whom there is a 16 percent or greater chance of being wrong. And, importantly, for those thirty-three students whose grades might easily be wrong, the error is large, one whole point on our 4.0 grading scale. For those wrongly graded students, the grade I would give says that they are a full standard deviation above or below average when they are not, or says that they are average when they are well above or below it.

Compare that to a system with finer gradations. If I divide up the range of all scores into fifteen equal subintervals of 5 points, from A+ through F-, all students are within one standard error of the border above and below their scores, and therefore I cannot be confident that I have placed any of the ninety-four students in the correct grade interval. However, if there is an error, the error is only one-third as large. Moreover, I am nearly certain that a student's grade is not in error by a whole grade point (3 standard errors), whereas errors that large were quite likely with the coarser gradations.

A finer grading scale increases the number of errors but reduces the size of the errors. Which should we choose in this tradeoff between frequency and

54. The standard error and other statistics are calculated by the scoring service at the Bureau of Evaluative Studies and Testing at Indiana University and reported to the teacher for each set of exams scored. The standard deviation of the scores is 15.4; the Spearman-Brown reliability of the exam is .91. There is a 68 percent chance that the true score is within one standard error of the measured score.

magnitude of errors? Both the theory of declining marginal utility of income and the actual and common sales of insurance suggest that the harm from putting large losses on a few persons is worse than the harm from placing small losses on many persons.[55] In addition, as a reader of undergraduate grades for purposes of law school admissions, I would much rather hear that all applicants' grades are slightly erroneous than hear that some of the applicants' grades are quite erroneous and others are correct, because the latter information prevents me from placing any weight on any of the grades. The information that the grades for every third student could be far off more deeply undermines confidence in all of the grades than the information that all are slightly off. For those reasons, reducing the gravity of errors by making grading intervals smaller and more numerous reduces both the unfairness and the inefficiency that come with unavoidable miscommunication.

It is true that drawing lines in the gaps can reduce the number of errors somewhat for either a fine or coarse scale, but doing so creates a different problem, lack of proportionality, which is discussed below. The better solution is a finer grading scale. But the adoption of a finer scale should not be accompanied by a requirement that all teachers use all of the points on the scale. What level of discrimination in assessment ought to be required of teachers is a separate issue implicating issues much broader than can be discussed here. If a teacher wants to give just six different grades, the adoption of a finer scale should not be the occasion for forcing her to expend more effort on grading or to shift to a method of assessment that she feels is inferior.[56] In fact, if she can see only A's, B's, and C's, when she reads the papers, it would be counterproductive to force her to randomly append a few pluses and minuses to those marks.[57] Conversely, one teacher's unwillingness to use a finely graduated system should not be sufficient reason for the school to fail to provide such a scale for those teachers who would find it equitable and useful for communication.

All of the points above can be accommodated without radically changing the way any teacher grades. I now turn to a principle that cannot be followed

55. It is my hunch that we should also prefer smaller, more frequent errors in the context of compiling GPAs. It seems less likely that errors in grading will average out when the errors are large and infrequent than when they are small and common.

56. A teacher who fails to discriminate between students who are meaningfully different sends messages that are both unfair and inefficient because she says that students are the same when they are not. But this is not the place for discussing whether she should change her method of writing or scoring exams. This article is about the process of turning raw scores into grades.

57. That is not to say that we should honor the teacher who wishes to round her finer numerical marks into rough groups on the ground that there is "no meaningful difference" between two different scores that she has awarded. To take an extreme example, suppose that the grading scale allows any score from .001 to .999. A teacher might say that there is no difference between her score of .343 and .344. Sometimes the teacher does not really mean "no difference"; the chances are greater that the .344 was better than the .343 than vice versa, however minuscule the difference. But if she says that there was truly no difference at all, then we should ask why she did not just use .34 for both and drop the random digit. If she also contends that there is no difference between .34 and .35, and then also between .3 and .4, then we ask how she can tell the difference between .4 and .5, or between .5 and .6, and so forth. Unless there is a discontinuity, she cannot tell the difference between the scores of .1 and .9, and should not be allowed to grade.

unless some teachers are willing to make at least small changes to their method of grading.

D. Use Proportional Grade Intervals

Grades are proportional, as I use the term here, when there is a linear mapping from the performance of the students, as assessed by the teacher, to the grading scale used by the school. For a teacher who uses a gestalt method, this means that the teacher assigns grades so that there is as much difference between an A paper and a B paper as there is between a B paper and a C paper. For a teacher who uses points, this means that the width of each grade interval is proportional to the difference between the numerical scores her school gives for the corresponding grades. Suppose a university has a grading scale in which A=4, B=3, and C=2, and minuses and pluses are .33 away from the whole grades. In other words, each possible grade in the university system is the same distance from the grades above and below it. A teacher's grading scale is proportional to the university scale if it is the same number of points from one grade to the next. The scale on which 17=A, 15=A-, 13=B+, 11=B, is proportional. By contrast, the scale on which 17=A, 15=A-, 12=B+, 8=B, is not proportional because the distance between the scores does not remain constant as it does in the university system into which the grades will be fed.

1. The Harm from Nonlinear Grading

There are two major ways in which lack of proportionality can lead to inaccurate communication. First, students with equivalent total performance can receive different grade point averages, and vice versa. Second, a reader comparing three students might wrongly infer that the difference in performance between an A+ and a A is the same as the difference between a B+ and B. I explain these two points below and then discuss a few reasons a teacher might be tempted to use disproportional grade intervals.

A disproportional grading scale will mislead readers of transcripts who harbor the reasonable presumption that a higher grade point average means higher total performance. Suppose all students take four courses given equal weight in the school's grading system. Although individual students perform differently in the courses, the students as a group perform essentially the same across the four courses; all of the exams have the same distribution of raw scores. Each teacher translates the raw scores onto a nonlinear scale employing the minimum cutoffs given below:[58]

Minimum raw score	*Grade*
40	4.0 (A)
32	3.5 (A-)
26	3.0 (B)
22	2.5 (B-)
20	2.0 (C)
18	1.5 (C-)
14	1.0 (D)
8	.5 (D-)
0	0 (F)

58. The teachers do not have to use the same nonlinear scale for the results herein to obtain.

Elba scores 19, 19, 36, 36 on her exams, and Ben scores 24, 24, 24, 24 on his. Ben deserves a 2.5 average, and he gets it. Elba also gets a 2.5 average, even though her total performance of 110 is substantially better than Ben's 96. Despite the fact that the teachers' grades have the same mean and standard deviation, two students with substantially different overall exam performance wind up with the same GPA. Of course, a similar error in communication can occur in any grading scale that rounds many scores onto a few grading points. But that is a separate problem of coarseness, already discussed. The inaccuracy here is not due to rounding because I chose the students' scores to be exactly in the middle of the various grade intervals. The use of a disproportional scale creates a communication error of its own in addition to the usual rounding problem.[59] When the relationship between the teacher's raw scores and the school's scale is not linear, ordinary addition can lead to poor signaling of comparative performance.

Lack of proportionality may lead to a more troubling outcome: different GPAs for students who performed the same. Suppose that Mark scores 21 on his four final exams, for a total of 84 points and a GPA of 2.0. Sara scores 16, 16, 16, and 36, for the same total of 84, but a GPA of only 1.625. Sara's total performance is the same but her GPA is worse. Except in unusual circumstances, we would not give a person who scores 21, 21, 21, and 21 on four equally weighted parts of a single exam a different grade from a person who scores 16, 16, 16, and 36. Yet that is just what we do to some students' GPAs when we fail to maintain proportionality in our grading scales. The inequity here will probably be exacerbated when the GPAs are converted to class ranks because that often has the effect of magnifying small differences in GPAs.

In the examples above, a student suffers for being less consistent in her exam performance. But that is not always the result. Suppose Kim scores 19, 19, 19, and 19 on her exams while Jake scores 21, 21, 21, and 11.[60] Kim's total of 76 is better than Jake's 74, but her GPA of 1.5 is worse than Jake's 1.625. We could also devise a nonlinear scale that would usually penalize consistency or one that would reward it. Unless a faculty decides that the more consistent performer should have a higher GPA, or the other way around, the school should use a scale that does not make consistency a hidden component of the GPA.

The unfairness of disproportionality in the case of Sara and Mark might be called horizontal inequity, in that essentially similar students are treated differently. Another way in which lack of proportionality can lead to unfairness might be called vertical inequity, which can crop up even when grades are not combined into GPAs and class ranks if a reader of the grades thinks that the interval scale is linear. Vertical inequity refers to situations in which different performances are treated either too differently or not differently enough. In a linear scale, the difference in underlying performance between a C and a B is the same as the difference in performance between a B and an A.

59. I see no reason to believe that the two errors tend to cancel each other in a way that might reach a second-best solution; in fact perhaps the opposite occurs.

60. Compare also 21, 21, 16, 16.

If a B student had scored *x* more points, she would have received an A; if she had scored *x* fewer points, she would have received a C. Under such a system, the performance deserving an F is as much worse than a C as an A is better. On a nonlinear scale, *x* fewer points might push a paper from a B to a D instead of just down to a C. In addition to unfairness, if grading is not linear, it is very difficult for a reader to later construct the meaning of an extreme grade since that meaning cannot be extrapolated from the difference in performance represented by various grades closer to the mean, the ones forming the bulk of the reader's experience.

The harm done by nonlinear transformations of raw scores into final grades might be brought home by considering the "grading" done by *U.S. News and World Report's* annual rankings of law schools. In the past, *U.S. News* turned most cardinal scores into ordinal scores. For example, Ivy School of Law might have a bar-pass rate of 89 percent, while Big Ten College of Law has a bar-pass rate of 91 percent. If 29 of the 180 schools had a bar-pass rate of 90 percent, Big Ten might have a rank of 40th while Ivy ranked 70th out of 180. This would make a mountain of impact out of a molehill of difference in what mattered, the actual rate at which students passed the bar.[61] As a result, schools' rankings could change from year to year even though the schools themselves had not changed in any meaningful way. Failing to maintain proportionality in the transformation of subscores led to inaccurate communication. *U.S. News* did not fully understand the disadvantages of its method when it published its early rankings.[62] It appears that the magazine has now corrected this error by using cardinal scores rather than converting them to ranks before combining them.[63]

2. Common Uses of Nonlinear Scales

There are good examples of and uses for nonlinear scales. The octave, the decibel, the pH scale, and the Richter scale are nonlinear. Yet three of those

61. At one time, *U.S. News* defended its rankings as being about the same as what schools do to students. See Mitchell Berger, Why the *U.S. News and World Report* Law School Rankings Are Both Useful and Important, 51 J. Legal Educ. 487, 499 (2001). Berger includes in his defense of the *U.S. News* rankings the statement that "no rankings system—including the grading system used in law school—is perfect." *Id.* at 500. Isn't it possible that there are important differences in the degree of imperfection? Although it is true that many teachers fail to maintain proportionality, I doubt that any law teachers (or schools) have violated this principle as egregiously as did *U.S. News* in its early rankings. It was particularly annoying to see *U.S. News* claim they are doing the same thing that we do, when they were clearly unaware of both how we do it and how to do it right.

62. The purpose of the ranking approach to the subscores was to prevent any large deviations on some criteria, like library size, from swamping other criteria. The better approach, as *U.S. News* now knows, is to use logarithms. For a discussion of its methodology, see Gayle Garrett, Our Method Explained, U.S. News & World Report, Apr. 15, 2002, at 55.

63. The page on "methodology" at the *U.S. News* Web site does not give a complete description of the method currently used, much less the methodology. See <http://www.usnews.com/usnews/edu/grad/rankings/about/03law_meth.htm> (last visited Feb. 10, 2003). The method in past years was not spelled out in detail either, but Joe Hoffmann and I figured out their method by working with the data and the results reached by *U.S. News*. I then called *U.S. News* to confirm that they had indeed done what Hoffmann and I had figured must have been done.

are defined by reference to a linear scale, and the other can be defined that way. If we are going to use a nonlinear scale, we ought to at least agree on its definition. To be consistent, any nonlinear scale will probably have to be defined in terms of a linear scale, because those are often more easily verifiable. In other words, if we define the scale nonlinearly, who can say whether the A interval is too far from the B interval to be consistent with another course? But there seems to be no reason to avoid a linear scale in favor of a nonlinear one. The usual reason to do so, that one extreme end of the scale is too far from the center, is not present in our grading of students. In law schools we rarely have some students scoring 10 times the score of other students, and never is one student's score 100 times that of another student.

One good reason to use a nonlinear grading scale would be to conform to the expectations of students or employers or admissions officers, that is, to communicate better with some audience. Is nonlinearity the usual assumption or expectation? Although there are many grading scales, perhaps the most common one makes 90 to 100 an A and 60 to 69 a D.[64] Based on my informal talks with students, I think that most college graduates, unless told otherwise, assume that the difference between an A and a B is the same as the difference between a B and a C, and so forth. Given the familiarity of linear scales, the burden would seem to be on those wanting to use a nonlinear scale to show that it better conforms to readers' expectations, or in some other way allows for better communication.

Another reason a teacher might wish to avoid giving different grades to closely spaced exams is that she is not confident her data mean that much. In other words, she likes to draw lines in the gaps. The benefit of doing this, as I've said, is that she does not falsely imply that two students close in actual performance (the bottom B and top B-) performed quite differently. (Note that this benefit decreases dramatically as the number of grades increases, i.e., when a finer grading scale is employed.) The undesirable flip side is that by stretching intervals to draw lines in the gaps a teacher also says that two students, the top and bottom B students, for example, are the same when they were significantly different. Or when she squeezes the B- interval to draw lines in the gaps, she says that the top C+ is further from the bottom B student than he really is. At best, drawing lines in the gaps tells a more accurate picture for some comparisons but a less accurate picture for other comparisons even within the same class, to say nothing of the difficulties it creates in making comparisons of GPAs. Drawing lines in the gaps carries no benefit large enough to justify the horizontal and vertical inequity that occurs when GPAs are calculated from grades given by teachers with disproportional grading scales.[65]

64. At the ends this scale is not linear, but scores in the nonlinear region are uncommon. When they do occur, however, the scale creates some problems. Take, for example, the student who gets 29 and 90 on two exams, each of which is "50 percent of the grade." The final grade, as calculated by some teachers, will be an F even though an A and an F would average to a C for other teachers.

65. The issue of skew was mentioned above. Since skewness is measured with a statistic that is based on all the performances in the class, it too could be standardized by schools with some beneficial effect. But doing so would often run afoul of this principle of maintaining proportionality.

Rankings of students are themselves nonlinear transformations of grades. When we publish class ranks, we create the possibility of error in communication. The 100th student is probably not very far behind the 90th in GPA, but the ranking disparity makes them look significantly different.[66] This problem might be cured by publishing the overall SPS for each student along with a distribution of the SPSs for all students. But even if a school chooses not to address the problems of communication stemming from its final ranking, it should attempt to prevent the problems of disproportionality at earlier stages in the process, for they compound the potential for miscommunication by changing not only the distance between students but also the actual order of the students in the class.

F. Keep the No-Credit Grade Reasonably Close to the Mean

A closely related principle is that the failing grade on the school's grading scale should not be too far from the mean grade. The fact that grades communicate relative performance does not mean that they cannot also communicate absolute performance. Many teachers have the sense that a certain level of performance is simply too poor to deserve credit for the course. Such a teacher might wish to give an F because he thinks the pupil (rarely a student) really deserves no credit for the course. If the grade that denies credit, usually the F, is far below the mean, the teacher's use of the grade implies that the student's performance is far below the mean.

At some schools the F is more than six standard deviations below the mean.[67] An F on such a scale implies that the student has performed much worse than the average. It is likely, however, that the failing student's performance was not *that* different from the mean. Forced to work within such a grading scale, teachers with high standards for what is passing work are put to an unfortunate choice between saying that the performance is passing when they think it is not and saying it is poorer than that of other students to a degree that is equally untrue.

There is an additional problem with the distant F. When the student's GPA is calculated, that F takes on much greater weight than it deserves. This is essentially a narrow instance of the point made above that a teacher who spreads his grades broadly gives his course more weight in the GPAs. The net result is that the teacher whose only intent is to say that the performance was

66. The unfairness of this rank-order tournament might be counterbalanced by the fact that rankings create large incentives for many, but not all, students to work hard.

67. On a normal distribution, which this is not, a random performance that far below the mean is about one case in a billion. Freund & Simon, *supra* note 21, at 523. Even the weirdest imaginable distribution will have few students more than 6 standard deviations from the mean. Chebyshev proved that the proportion of the data that lie within k standard deviations of the mean is at least $(1-1/k^2)$. Hence, the number of observations outside of 6 standard deviations is at most 1/36 for any possible distribution. For nearly all sets of data, the actual percentage of data beyond 6 Σ is much less than that. See *id.* at 78–79. In the 2,000 exams I have graded with a point system in the last 10 years, I have not seen a single one. Thus, even accounting for the fact that our distributions may not be normal for the reasons given above, a grade 6 Σ short of the mean ought to be rare.

not good enough for credit may inadvertently push that failing student a long way down in class rank, much further down than that student deserved to be pushed by a poor performance in a single course. To avoid this, teachers should have available a grade that tells the reader that the performance did not deserve credit without implying that the student performed so far from the mean that we expect to see such performance only once in a blue moon.[68]

There are some contexts in which teachers may be especially likely to apply fixed rather than relative criteria in assigning grades. If the faculty has decided, for example, that certain courses must be passed in order for a student to graduate, teachers of those courses should probably feel that they have been given a special obligation to determine whether the students did indeed learn enough of the subject matter to have cleared that hurdle. Especially if the course is mandatory and cannot be repeated, if there are any such courses, the requirement would seem to assume that some sort of absolute standard ought to be applied in grading.[69] If the failing grade is too far from the mean, the responsible teacher of a low-scoring student is in a difficult bind.[70]

G. Avoid Grade Inflation

Once we have established the principles above, it becomes obvious that grade inflation generates a number of problems. When some teachers increase grades, they create incentives for students to take their courses rather than other courses. At the margin, the higher grades in a grade-inflated course will cause some students to take that course instead of a course that would have been better for their educational development. That in turn puts pressure on other teachers to inflate their grades because many teachers want to have students in their elective classes. With time, there is pressure on other schools to inflate their grades so that their graduates have a fair shot in the competitive education and employment markets. Thus, grade inflation spirals.

One result of inflation is that grades are harder to compare over time, although such comparisons might not be terribly meaningful for other reasons.[71] In addition, grades are harder for the readers to interpret because they

68. The discussion has assumed that credit for a course will be tied to grades in some way. Another way to solve the problem would be to have teachers determine two separate issues, the grade and whether the student receives credit for the course.

69. If a diploma carries with it the right to practice law in a state, it might be appropriate for the teachers in schools granting such diplomas to calculate their grades according to some fixed standard. The same point could be extended to all law schools on the theory that we are one hurdle that must be crossed to gain admission to the bar in most states, but the vast differences in students across schools (not to mention the teaching) make it nearly impossible for any of us to have confidence in our ability to set an appropriate standard based on who should be allowed into the bar.

70. The teacher of a pass/fail student might be especially troubled. It is possible to argue that the level of performance required for a pass/fail student to pass should be higher than the level required for a regular student to pass. If a different standard is not applied, the pass/fail student will have almost no incentive to study. If a faculty takes that position, then teachers become more likely to award F's, and the resulting F's are quite misleading as signals of the relative performance of the students.

71. Teachers and students both change over time. For example, most schools have more competitive admission requirements than they had in the past.

do not know how much inflation has occurred. People who read lots of transcripts from a particular institution should have no trouble keeping up with the shifting meanings of the grading language, but others will not know how to interpret the message.[72] As meanings shift at different rates within institutions, it becomes even harder to make comparisons of students across institutions.

Another problem created by grade inflation is that it reduces the number of useful grade intervals, turning fine grading scales into coarse grading scales. As was shown above, coarser grading scales create serious problems of injustice in addition to ambiguity in communication.[73] Still another problem created by inflation is that it warps proportional grading scales, with all of the concomitant problems of unfairness and inefficiency in communication noted in the discussion of disproportional scales. The absence of proportionality was probably not as large a problem back when C was the average and an F need not have been much rarer than an A. Teachers in those days had the elbow room to fail students when they fell a couple of standard deviations below the average.

Ironically, grade inflation can be especially harmful to students on the bottom rungs of the class-rank ladder. When the mean moves up, it puts the teacher of weak students in a bind. He has two choices, retain linearity and give no failing grades, or abandon linearity. If he has a strong sense of what is required for passing performance, he will refuse to move the bottom grades up into the passing intervals, even though that is where the bottom students belong if proportionality is to be maintained. Grade inflation has elevated and compressed the scale so much that, as I've said, the failing grade at IU Law is now about six standard deviations from the mean. The chances that a student's true score was that far from the mean are minuscule. When B is the average and the standard deviation is half the distance from B to A, an F should be exceptionally rare.

Schools might respond to grade inflation by raising the average required for graduation. Unfortunately, not all of them have similarly raised the grade required for passing performance. The no-credit interval has been the only portion of the scale not to inflate, stretching the whole scale beyond sensibility. This has dramatically shifted the role of the teacher. Instead of being asked to determine whether a student gets credit for the course, the teacher is now asked to determine whether it will be difficult for the student to retain his scholarship without taking a disproportionate number of courses from easy graders. In some cases the F may even make it hard for the student to graduate. It is doubtful that the teacher of a single course has enough data to justify giving a grade that prevents a student from continuing in school. The teacher grading her own students should not have to shoulder the burden of justifying such a drastic result when all the teacher wanted to do was say No Credit for her course.

72. The fact that almost 90 percent of Harvard's graduates qualified for "honors" must have come as a surprise to at least a few uninformed observers. See Anemona Hartocollis, Harvard Faculty Votes to Put the Excellence Back in the A, N.Y. Times, May 22, 2002, at A20.

73. As grades distinguish less, employers might be inclined to shift their reliance to exams that do distinguish, but do so perhaps even less reliably, such as the multistate bar exam.

It is possible that some teachers offer courses with a high minimum grade because they perceive that the abnormally low grades given in other courses have more impact on the GPA than they should have. Thus another vicious circle is born. Rather than taking that vigilante approach, we should correct defective grading scales by increasing the minimum grade required for credit. This would not be hard to do. GPAs could continue to be calculated as they are now, but credits completed toward graduation would not include any courses in which the student received a grade lower than C (for example). If we were to do that, responsible teachers wanting to say No Credit could do so without giving students grades that should be statistically impossible.

I am not making any argument about how much grade inflation there has been—an issue on which there has been some debate on our campus. Nor have I made any attempt to marshal the reasons grade inflation might be appropriate, such as the fact that at many institutions the students admitted are better qualified than in the past, or that peer schools have inflated their grades so much that it would be miscommunication not to inflate the grades in a laggard school. (Another way to deal with the latter problem is to adopt an unusual scale.) In other words, I am not trying to argue that a school should squeeze all of the inflation out of the grading scale any more than the Federal Reserve should try to squeeze all of the inflation out of the economy. I am arguing, however, that grade inflation leads to miscommunication, which can be both unfair and inefficient. If a school allows inflation, it should take steps to minimize that potential.

* * * * *

Any time a faculty imposes grading constraints on its members, it risks forcing miscommunication. If the class is abnormal, the constrained teacher is unable to send an accurate message of comparative performance because she cannot award abnormal grades. On the other hand, if the class is normal, an unconstrained teacher with a defective assessment instrument might send inaccurate messages of comparative performance by sending abnormal grades. One key question, the question some faculties have failed to ask, is which risk is greater. Is it more likely that the teacher's instrument is miscalibrated, or that the class itself is abnormal? For classes of many students, the answer will usually be that it is more likely that the assessment instrument is defective. For small classes, the risks become closer. Because these risks are somewhat speculative, increased uniformity in grading will not necessarily lead to increased accuracy in communication. Indeed, grading constraints will undoubtedly lead to some sets of grades that are less accurate than they would have been if the teacher had been unconstrained. But the issue is not whether uniformity will improve matters in every case, but rather whether it improves communication in the long run. Forced uniformity will often increase accuracy, and overall the odds lie against complete professorial freedom in grading.

I have spoken of uniformity as if it must be forced upon the faculty because there are incentives for teachers to deviate from norms if they are allowed to do so. Nevertheless, a very collegial faculty might accomplish as much with informal norms as would be accomplished by binding rules. And even if

informal norms or formal rules are flouted by some, partial adherence has some beneficial effect. Generally, the more adherence to the principles above the better, regardless of the source or formality of the constraint.

Assessment of student performance is never perfect. Choosing the types of instruments for measuring performance will always involve tradeoffs. Nevertheless, we should try to avoid exacerbating deficiencies in our assessment by making errors in the numerical methods we use for coming to summary comparative statistics. Judgments about proper grading practices must take serious account of the context in which our grades reside. If we are to give grades and class ranks, let them be fairly calculated. We teach that justice matters; let us do our best when it is our turn to hand down the decisions.