Grading, class selection, and work in theory and practice at Amherst[1]

Submitted to the Department of Economics of Amherst College in partial fulfillment of the requirements for the degree of Bachelor of Arts with honors.

Faculty advisor: Geoffrey Woglom
Faculty readers: Jessica Wolpaw-Reyes and Daniel Barbezat

David Gottlieb

May 4, 2006

[1]I would like to thank the whole faculty of Economics at Amherst College. In particular, I want to thank Prof. Geoffrey Woglom, my thesis advisor as well as my major advisor, for all of his help with this project and for always taking a hard look at my class schedule. I'm also indebted to him and to Prof. Steve Rivkin for hiring me to work on related projects in summer 2005, and for helping me get my hands on the dataset used in this project, and to Prof. Rivkin in particular for forgiving me for not solving his robust standard error problem. Thanks to my faculty readers, Profs. Daniel Barbezat and Jessica Wolpaw-Reyes. I also want to thank Dean of the Faculty Greg Call for approving the use of student data in this project, and for teaching me Intro to Analysis freshman year. Thanks to Min Wang for all of her caring and support. Thanks to Sarang "Opal Mehta" Gopalakrishnan for sometimes being messier than me when we were roommates. Thanks to my parents, brother, and dog for sending me to college. And thanks to Daisuke "Dice" O for getting a job.

Abstract

We undertake a theoretical and empirical analysis of the problems involved in grading generally and, in particular, in grading at Amherst College. In the theoretical portion of the paper, we develop a model of grading that takes into account two particular (and unusual) features: granularity and noisiness. We analyze grades first in terms of their information content, and then look at their incentive effects. Since less-granular grading systems in effect produce higher grades, *ceteris paribus*, students will tend to choose classes with such grading systems. The same effect also influences incentives to work: students will generally work harder when the grading system is more fine-grained. Interestingly, the amount of noise in a grading system influences students' incentive to work. In the presence of noise, students may work harder or less hard depending on risk aversion. Finally, as anticipated, we find the amount of work a student expects to do in a given class will influence class selection. In the empirical portion we find that, from class to class and department to department, there's a lot of apparent variation among grading systems, both in their granularity and their noisiness. When we apply our theoretical tools, a somewhat disheartening picture emerges. We consider the implicit policy questions, and review solutions: how can we fix grading, preferably without introducing a fascistic "command and control" regime?

Contents

Chapter 1

Introduction

The subject of grading should intrigue any economist. Grades in school have a lot in common with our favorite thing (money) – grades send behavior-influencing signals just like prices, grade inflation is almost as terrifying as regular inflation – but there are important differences, too. Giving out grades isn't a zero-sum game in the same way as passing out dollar bills is, so the cost of giving out high grades may be lower than it ought to be. And, unlike the wage scale, the grade scale is capped above, which means that what's commonly called "grade inflation" is really more like "grade compression" and raises a whole slew of interesting questions that don't come up in the context of monetary inflation. Despite their intrinsic interest, though, such features are dealt with only piecemeal and incompletely in the literature. Drawing some insights from the published literature, we'll develop a general model of grading, class work, and class selection that allows us to evaluate the consequences of several types of variation in the way classes are graded. Classes where finer distinctions of grade are possible will tend to induce students to work closer to what seems like a desirable amount. Classes with more "noise" – that is, classes where grades are less dependent on ability and effort – will produce anomalous decisions and behaviors whose exact character depends on risk tolerance.

Looking at the Amherst College data, we'll find the theoretical lessons we've derived have a lot of relevance. The evidence will strongly suggest that classes and departments vary widely in how fine-grained their grading systems are, and also in how noisy their grades are, which means that both class selection and the information content of grades may be badly broken. The policy question – what to do about inter-class and interdepartmental grading differences, if anything – is already the subject of a fairly extensive literature. Correctly eschewing fascistic solutions like directly controlling how professors grade students, several economists (and a couple schools) have espoused grade post-processing schemes, whereby students' transcripts and GPAs are subjected to weighting mechanisms of varying complexity to produce adjusted grades that account for differences in how classes are graded. There are now so many such schemes that it's not immediately obvious which are the best, but our theoretical framework will point the way toward making more enlightened choices in the future.

Our discussion of grades will be streamlined if we can provide general answers to a few important questions before our exposition begins in earnest. In particular,

- What should grades do?
- Is there grade inflation?

- What are the consequences of grade inflation?

1.1 What should grades do?

The published literature doesn't furnish a fully general, over-arching theory of what grades do, but more modest conclusions abound. In general, they fall into two categories: full discussions in theoretical papers about what grades should do, and brief mentions in empirical papers about how well grades work in practice and how they might be fixed. The former kind of paper seems to be somewhat rare, probably because the purpose of grades is usually taken for granted, but there are examples, like [Dubey and Geanakoplos2004] and [Ostrovsky and Schwarz2002]. These papers are notable (and, likely, were written) primarily because they challenge the conventional wisdom on grading; both argue that disclosing less-than-perfect information through grades might be optimal. Dubey et al. argue that, given certain assumptions, students might respond more desirably to the incentives of general rank grades (As and Bs) than of specific number grades, and Ostrovsky et al. argue that transcripts which reveal only partial ranking information likely represent an optimal (or at least equilibrium) compromise between the various players (students, universities, and employers). As is often the case, though, the conventional wisdom has a lot of merit, and while such ideas are interesting, I think of them as exceptions that prove the general rule that getting better information from grades is better, both because of students' incentives and for the convenience of later transcript readers (e.g. employers, awards committees, etc.). As mentioned, this is the consensus view, and usually briefly conjured in empirical papers on related subjects. For example, [Johnson1997] posits that when two students take the same class, the better student usually receives the higher grade, and an ideal grading system should (statistically speaking) give the better student a higher average across all his classes [Johnson1997, 260]; his Bayesian model [Johnson1997, 255] implicitly incorporates this. Most authors of such papers seem to be trying to compute an adjusted grade that measures "student quality" (usually a one-dimensional property, with class, subject, etc. variations captured in error terms), which implicitly suggests the prescription that grades should measure performance or achievement in a way that's uniform across all classes. Varied justifications are presented for this belief, and more than one of them is probably correct and significant. For example, [Sabot and Wakeman-Linn1991] argue that grades should serve as signals to students of their comparative advantages across various disciplines, to encourage more efficient specialization (160, e.g.). Obviously, more accurate grades are also more useful from a signaling perspective – employers can decipher transcripts more easily if students do not effectively receive arbitrary GPA bonuses based on class selection.

It's easy to make a couple prescriptions that represent an overview of this literature: the better student should receive the better GPA, and students should receive better grades in the subjects in which they are advantaged. Relative grade inflation interferes with both of these.

1.2 Is there grade inflation?

Yes.

Grade inflation has been recognized for decades. Absolute grade inflation (wherein all classes have uniformly inflated grades) has been documented. Relative grade inflation is also a known phenomenon, and has been raised as an explanation for students' shift away from the sciences and towards the humanities ([Sabot and Wakeman-Linn1991] contains a discussion of related publications, as well as evidence from Williams College).

1.3 What are the consequences of grade inflation?

Absolute grade inflation is problematic because it narrows the range of available grades, which makes grades less precise and comparisons between students less fine-grained and more error-prone (though the importance of this has been disputed; see [Ellenberg2002]. Relative grade inflation is even more problematic, of course. [Johnson1997] summarizes the problems fairly well: students will tend to sort into the classes and / or departments that are graded more easily, professors face pressure to inflate their own classes' grades to compete for students and receive high marks in evaluations by students, and this pressure tends to flatten grades, which diminishes the rewards that can be given to the most excellent students [Johnson1997, 266].

Chapter 2

Theory

Word of mouth and personal experience suggest that departments vary in the grading standards they apply. In some classes, it's quite hard to get an A, and in some it's hard not to; in some a wide range of grades are possible, while in others everyone seems to get the same grade. Further, in some classes, brilliance is obviously rewarded richly, whereas in others it's often unclear what's being rewarded. We can extrapolate some broader conclusions about the way grades are assigned in different classes. For example, in a department where grades are both higher and less widely dispersed, it seems like a smaller grade scale is employed – a scale that runs from A to C rather than A to F, for example – or, at least, that most of the possible range of performance in the class is covered by a smaller grade scale (so that a student who never attends class may fail, but students who always come to class are effectively graded A to B). Similarly, a department that appears to offer smaller rewards to students' generalized academic ability is giving some other factor(s) relatively more weight in grading, whether meaningful (e.g. effort, interest) or random.

In this section we'll develop a simple model of grading and class performance to test what effects such various grading structures might have on class performance, learning, and class selection. We'll formalize the observations above. In particular, what effect do higher average grades (and the accompanying diminished grade range) have on the informational content of grades (the relative shares of signal and noise)? Will students choose courses with higher average grades to improve their own grades (answer: yes; see [Sabot and Wakeman-Linn1991])? How do factors like the granularity (detail level) of the grade scale and the returns to ability affect students' incentive to work? How do all of these play into expectations about work load that affect course selection? Finally, we'll draw back a little and discuss how these findings play into the overall character of the educational institution, and how professors' grading incentives might be influenced.

2.1 Definitions

Performance interval For the jth class, the performance interval $[0, \overline{p_j}]$ defines the gamut of possible performances for any student in the class. Every student's overall performance in the class corresponds to a point in the interval, with 0 representing the worst possible performance and $\overline{p_j}$ representing the best possible performance[1].

[1] This implies we're assuming the range of possible performances in a given class is bounded above and below, but we could actually relax either or both of those assumptions without materially changing the exposition. If we want to assume instead that performance above $\overline{p_j}$ is possible, we need only require that it be graded identically to performance equal to $\overline{p_j}$ – an easy assumption, since real-world grade scales actually are bounded above. Given that the boundedness assumption is not strictly necessary, we'll relax it as

Performance function For the ith student and jth class, the performance function is

$$P_{i,j} = A_i^{\alpha_j} E_{i,j}^{\beta_j} + \epsilon_{i,j};\ \epsilon_{i,j} \sim N(0, \sigma_j), A_i \geq 0, E_{i,j} \geq 0;\ 0 < \alpha_j, \beta_j < 1 \tag{2.1}$$

where A_i represents student i's generalized academic ability, $E_{i,j}$ represents the effort student i invests in class j, and $\epsilon_{i,j}$ represents the noise present in generating and / or measuring performance in class j. The multiplicative nature of this specification also captures the intuition that ability (or effort) alone exhibits diminishing returns to scale (and ability and effort are therefore complementary in performance). There are some limitations to this approach. For example, it only explicitly accounts for the set of abilities that all classes draw upon in common. However, this doesn't seriously diminish the validity even of analyses that treat classes that actually do draw upon somewhat differing abilities – if performance in a fine arts class J depends on a student's general academic ability, A_i, and upon artistic talent, F_i, and F is distributed independently of A, then we may treat F as a component of the noise variable, $\epsilon_{i,j}$. The definition of performance works properly as long as it represents performance as a function of ability, effort, and noise.

Grade system A grade system is a subdivision Γ_j that divides the jth class's performance interval $[0, \overline{p_j}]$ into one or more subintervals of equal size. The subintervals are labeled A, B, C, etc. from highest to

it's convenient.

lowest, each representing a possible grade. So if Γ_j divides $[0, \overline{p_j}]$ into two subintervals, $[0, \overline{p_j}/2]$ and $[\overline{p_j}/2, \overline{p_j}]$, then a student whose performance is (*e.g.*) $\frac{1}{4}\overline{p_j}$ will receive a B, and a student whose performance is $\frac{3}{4}\overline{p_j}$ will receive an A [2].

Grade function The grade function of the ith student in the jth class is

$$G_{i,j} = \begin{cases} 4 \text{ if } P_{i,j} \text{ is in interval labeled A} \\ 3 \text{ if } P_{i,j} \text{ is in interval labeled B} \\ 2 \text{ if } P_{i,j} \text{ is in interval labeled C} \\ 1 \text{ if } P_{i,j} \text{ is in interval labeled D} \\ 0 \text{ if } P_{i,j} \text{ is in interval labeled F} \end{cases} \tag{2.2}$$

This definition captures the intuition that grades can be ranked straightforwardly and sequentially, and will simplify construction of a utility function that accounts for grade differences. Obviously a somewhat different function would be needed for grade systems with more than 5 subintervals, but such an extension is neither difficult nor important. We should also note that this kind of grade function also in effect incorporates the assumption that it's impossible to distinguish an A in a given class from an A in another class (even if they vary in difficulty), since it assigns them the same numerical value. In real terms, this is an assumption about the information available to those

[2] Equally sized subintervals is a neat simplifying assumption, but has limitations. Most notably, if we allow the subintervals to vary in size, we can capture the grade system of a class where it's technically possible to get any letter grade, but almost all students get As or Bs: the subintervals representing C, D and F are small and packed tightly close to 0. Fortunately, relaxing the equal-size assumption would not substantially effect our exposition.

who read students' transcripts; it assumes that they don't have the information to distinguish one class from another. If we wanted to make the polar opposite (optimistic) assumption about course information available to transcript readers, we would structure the grade function to incorporate that information. In our model, information about how a course is graded translates into the details of its grading system and performance interval. So if class j has performance interval $[0, \overline{p_j}]$ and grading system Γ_j, we'll incorporate the information about what range of performance a particular grade corresponds to and write

$$G_{i,j} = \text{ the midpoint of the subinterval of } \Gamma_j \text{ that contains } P_{i,j}. \tag{2.3}$$

Perfect grading Throughout our treatment of the deeply imperfect grading systems that populate the world, we'll want an idealized standard to refer back to. Our "perfect" grading system will incorporate two special features:

1. The "optimistic" assumption about course information available to transcript readers (as detailed above).
2. Continuity. Instead of dividing the gamut of performance into a number of discrete intervals, our perfect grading system will cover it smoothly.

It's actually quite easy to devise a grade function formula that meets these requirements: just set grade equal to performance.

$$G_{i,j} = P_{i,j} = A_i^{\alpha_j} E_{i,j}^{\beta_j} + \epsilon_{i,j} \tag{2.4}$$

> Why call this system "perfect"? Well, without a complete analysis of the costs and benefits of college education we can't make any final conclusions, but this system has qualities that seem intuitively quite attractive: it matches student accomplishment (performance) to student reward (grade), and it transmits the finest-grained performance information to transcript readers.

As we've constructed the model, every class has its own performance interval, and its own grade system laid over that; these features define students' possible performance. A student's grade is determined by which subinterval of the grade system his performance function lands in.

2.2 Information content

The most obvious purpose of grades is to convey information about performance, to students, advisors, grad schools, employers, and mom and dad. In particular, grades are supposed to convey information about ability and effort to whomever is reading the transcript. We'll make the optimistic assumption about transcript readers' information (see 2.1)[3]. Someone reading student i's transcript and looking at a grade G in class j will make an

[3]Naturally the quality of information is worse if the pessimistic assumption is true and grade information comes from more than one class.

inference like the following:

$$(A_i^{\alpha_j} E_{i,j}^{\beta_j}) \in \text{ the grade system subinterval labeled with grade G.} \qquad (2.5)$$

This inference can be correct or incorrect (since there's a noise component to the true performance function), and it can also be more or less meaningful (in a grade system with only one subinterval, it's meaningless; with two subintervals, it's more meaningful; with three subintervals, it's even more meaningful)[4].

Broadly, we can discern two aspects of a grading system's informational content that are interesting: accuracy (that is, the frequency with which transcript-readers' inferences will be correct) and precision (how specific the inferences are). So long as there's noise in the performance function, the two trade off in the obvious way: *ceteris paribus*, more accuracy means less precision and more precision means less accuracy. A grade system with just one subinterval is perfectly accurate, since transcript-readers' inferences regarding which subinterval performance falls into will always be correct, but also perfectly imprecise – a grade from it conveys no performance information. Likewise, for any particular student's performance in a particular class (with noise), $\exists N \geq 1$ such that if $n \geq N$ a grading system with n subintervals will always misgrade that student (of course this does not mean such a grading system has zero accuracy), but the general rule is that as the number of subintervals in the grading system increases, precision increases and

[4] The incentive purpose of grades can be thought of as an extension of the informational; grades should match the outward results students want to achieve with the performance employers will believe they reflect.

accuracy diminishes. The optimal grading system in terms of information content is indeterminate unless we assign specific relative values to accuracy and precision, but intuitively it seems like an arbitrarily precise grade system would be desirable; transcript-readers can account for a small possibility of error, but they can't impute more detail than is presented in the transcript. Further, the total amount of inaccuracy is constrained by the magnitude of the noise variable, so as the grade system becomes arbitrarily precise, the inaccuracies tail off.

Another way of thinking about accuracy and precision may help us here: suppose transcript-readers look at grades mostly to determine students' relative ranks rather than absolute achievement. For a given grading system and two students, then, three meaningfully distinct scenarios are possible:

1. The system grades (and thereby ranks) the higher-performing student better than the lower-performing student.

2. The system grades the two students the same.

3. The system grades the lower-performing student better than the higher-performing student.

A more precise grading system is one in which scenario 1. occurs more often; a more accurate grading system is one in which scenario 3. occurs less often. For class j, as the number of subintervals in the grading system increases, the probability of scenario 2. goes to zero and the respective probabilities of scenarios 1. and 3. converge on limiting values (determined by σ_j, the noise in the performance function of the class – for large σ_j, the limiting probability of scenario 3. will be large; for small σ_j, the limiting probability of scenario 3. will be small[5]).

This discussion points us to a distinction between two possible types of deviation among students' grades we may observe empirically: scenario 1. (good) deviation, and scenario 3. (bad) deviation. In class j, good deviation is caused by differences in student quality (A_i) and effort ($E_{i,j}$) (and by precision in grading) and bad deviation is caused by noise (σ_j) in the performance function. Distinguishing between the two in theory is easy: if A_i and $E_{i,j}$ are known for all i, good deviation is the amount explained by variation in those variables; bad deviation is everything else. The difficulty arises when trying to distinguish between the two in practice; our ability to distinguish good from bad deviation depends on our ability to estimate ability and effort, which is tricky. The empirical section contains several attempts at this.

The over-arching problem with grades' information content may be that professors don't have a strong incentive to produce grades that carry good information. They face no shortage of high grades, after all, and if they want to make a particular student really stand out they can just write a glowing recommendation.

[5] This supports our earlier intuition that an arbitrarily precise grade system would be desirable (if achievable) – the inaccuracies approach a limiting value and, if σ_j is sufficiently small relative to the population deviation in ability and effort, the probability of scenario 1. will grow at a higher rate than that of scenario 3.

2.3 Course selection and maximizing grade

It's intuitively plausible that students take grading into account when they select their courseloads, although naturally other things are considered, too, like expected workload (which we'll deal with later), interest in the course material, peer preferences, prerequisites of desired courses, etc. As noted, this problem has already been considered at our rival college (see [Sabot and Wakeman-Linn1991]); we will find that although our respective student populations are not easily commensurable (and ours vastly superior), a similar logic is likely to motivate students at every school. In this section we construct a basic theoretical framework that validates that basic intuition, and lay the groundwork for the slightly more complicated problems to come.

Students making course choices are like any agents making any choices – they require some way to rank available options against each other. We will furnish an explicit utility function and assume students make the choice that maximizes their utility; this approach will serve us in later sections as well. We'll give student i the utility function

$$U_{i,j} = G_{i,j}^{\gamma_i} - E_{i,j}^{\delta_i};\ 0 < \gamma_i, \delta_i < 1, \tag{2.6}$$

with $G_{i,j}$ and $E_{i,j}$ signifying the student's grade (per grade function) and effort in the class taken, respectively[6]. Ignoring effort for now, it's easy to

[6]Some factors are not considered. For example, a student will find some classes more interesting than others, and derive utility from enrolling in them independent of other factors. Likewise, relative rank in class may count for something as well. However, the empirical work in [Sabot and Wakeman-Linn1991, 167] specifically supports the position that absolute grade level (not just relative rank in class) matters a lot in this determination,

see how grade differences between classes could influence choice. If student i will take either class 1 or class 2 (whichever yields higher utility) and $G_{i,1} = 4$ (an A) while $G_{i,2} = 3$ (a B), if $\gamma_i > 0$ (which is probably true for all students), then $U_{i,2} > U_{i,1}$ and the student will choose class 2. This result is not tremendously interesting in itself. However, the methodology is useful: we needed a general framework to evaluate students' class selection decisions, and now we have it.

2.4 Grading and incentives to work

Grades are a powerful motivator; short of threatening violence or cutting off food and water (both illegal), grades may be the most effective means professors have for manipulating student behavior. We've already seen the dark side of this effect: raising all the grades in a class will probably increase enrollment in that class even when it's not socially optimal to do so. But grades *are* socially useful. In course selection, grades can serve as signals of comparative advantage [Sabot and Wakeman-Linn1991, 164] (as Sabot et al. note, this purpose is served best when different academic disciplines all grade according to similar grade systems). Further, grade incentives affect how much work students do in the classes they've chosen. Few would make sacrifices for their studies in any classes if their grades didn't depend on it. In fact, we will find grades' influence enter students' work decisions in multiple ways.

---

i.e. γ_j is large compared to the coefficients to other factors.

2.4.1 The grading system and incentive to work

Obviously the width of a class's grade scale is one factor that will affect students' desire to exert themselves. If there's a wide range of possible achievement in a class, there's more payoff to traversing that range. Consider the trivial grading system Γ_1 with only one subinterval. If class j has system Γ_1, only one grade is possible, so a student's effort can never improve his grade. Students would do little work in such a class (as our model indeed confirms).

On the other hand, consider class j with perfect grading (see p. 11). Let's analyze the effort decision[7]:

$$U_{i,j} = (P_{i,j})^{\gamma_i} - E_{i,j}^{\delta_i} \tag{2.7}$$

$$U_{i,j} = (A_i^{\alpha_j} E_{i,j}^{\beta_j} + \epsilon_{i,j})^{\gamma_i} - E_{i,j}^{\delta_i}, \tag{2.8}$$

and assuming there's no noise in the performance function (convenient for now), the first-order condition is

$$E_{i,j} = (\frac{A_i^{\alpha_j \gamma_i} \beta_j \gamma_i}{\delta_i})^{\frac{1}{\delta_i - \beta_j \gamma_i}}, \tag{2.9}$$

or, in terms of marginal benefit and marginal cost:

$$MB = A_i^{\alpha_j \gamma_i} \beta_j \gamma_i E_{i,j}^{\beta_j \gamma_i - 1} \tag{2.10}$$

$$MC = \delta_i E_{i,j}^{\delta_i - 1} \tag{2.11}$$

[7]Assuming here that performance intervals are not bounded above simplifies our analysis without substantially affecting the result. The only difference is that if performance intervals are bounded above (as in reality they are), some students will only work until they achieve the highest possible grade.

$$\text{FOC: } MB = MC. \tag{2.12}$$

Students decide how much to work based on their personal preferences for work and achievement, and on how much they'll gain from exerting themselves. Optimal effort is increasing in ability (since they're complements), γ_i, and α_j and β_j; it's decreasing in δ_i. All of this is both unsurprising and normatively correct.

Of course, real grading systems can't be perfect. Real grading systems are modeled by Γ_j with finitely many subintervals, which means the level of detail (and thereby responsiveness) given above (equation 2.9) are not achievable in real life. However, it makes sense to think of the perfect grading system as the asymptotic case which real grading systems approach as their number of subintervals increases (a grading system with a ton of subintervals might as well be continuous). Classes with more variation in grades correspond to grading systems with more subintervals, and the incentives in those classes will more closely approximate those given in the perfect case above. Since we must deal with imperfection, though, we should consider what the particular character of the imperfections introduced by grading systems with finitely many subintervals will be[8].

First consider the general impact of an imperfect grading system (and discontinuous grade function, as on p. 10). Most basically, the marginal benefit is quantized rather than continuous – instead of a student's grade improving gradually as he expends more effort, in a class where the only grades are A and B a student whose ability puts him in B territory will see

[8] We'll preserve the "no noise" assumption for now.

no payoff to effort unless he works hard enough to get an A. The marginal cost, however, (the disutility caused by effort) is still continuous. So the graph of utility by effort will be a series of downward-sloping plateaux, each representing a grade level. The highest point on each plateau is the leftmost, that is, the least-effort point for given grade. In other words, in an imperfect grading system, students will do exactly enough work to achieve their desired grade and *no more*. This is an important, if intuitive, result: an imperfect, quantized grading system will tend to diminish the amount of work students do relative to the perfect grading system (the presumptive social optimum).

This points to a real-world trend: grading systems that have more subintervals (approximating perfect grading more closely) will generally give students incentive to work harder, or at least closer to the socially desirable "perfect" amount[9].

The basic conclusions of this section are that

- Finer-grained grading systems induce students to work harder; coarser-grained grading systems allow students to slack once they hit their target grades.
- Finer-grained grading systems are better in that they induce students to approximate more closely the "natural" level of effort they would exert if the payoff to them was based on true performance rather than grade.

[9] Obviously we haven't proved this proposition in its full generality, but it's especially obvious when we think of subdivisions that refine other, less fine subdivisions.

2.4.2 Performance noise and incentive to work

To this point, we've been studiously ignoring the random noise built into the performance function. However, this too can have profound consequences for student behavior, and so the amount of noise in a class is a factor we should take into account when thinking about how students will respond to that class. Consider the following: Yojimbo is taking a class, Quantum Field Theory, in which two possible grades (A and B) divide a performance interval of $[0, 1]$ such that performance $P_{i,QFT} = \frac{1}{2}$ or higher earns an A and anything else earns a B. Assume the performance function

$$P_{i,QFT} = A_i^{\frac{1}{2}} E_{i,QFT}^{\frac{1}{2}} \tag{2.13}$$

Suppose Yojimbo's ability $A_Y = \frac{1}{4}$. Assume

$$U_{Y,QFT} = G_{Y,QFT}^{\frac{1}{2}} - E_{Y,QFT}^{\frac{1}{2}} \tag{2.14}$$

so it's utility-maximizing for Yojimbo to get an A. Then he must work at least $E_{Y,QFT} = 1$. Our analysis above shows he will work exactly that much and no more, and, as above, this may be a social loss, in the sense that if Yojimbo's reward were more closely related to his production he would work more. Above, we argued that refining the grading system could increase Yojimbo's incentive to work and improve the situation, but suppose we can't do that – the professor doesn't think bright kids deserve grades lower than B, maybe, or there are too many students and too few teachers to increase the level of detail in grading. What else could increase Yojimbo's incentive to work and push him closer to the optimal level of effort? How about throwing a little scare into him? Let's introduce some uncertainty; if Yojimbo doesn't know the precise value of his performance function, he's going to be much less confident about fine-tuning it so it lands at $\frac{1}{2}$ and maybe he'll work extra just to be safe.

To simplify the exposition, let's use the noisy performance function defined above, but substitute a uniformly distributed noise variable for the normal one supposed earlier[10].

$$P_{i,QFT} = A_i^{\frac{1}{2}} E_{i,QFT}^{\frac{1}{2}} + \epsilon i, QFT, \ \epsilon_{i,QFT} \sim \text{unif. over } [-\frac{1}{4}, \frac{1}{4}] \tag{2.15}$$

Consider Yojimbo's optimal effort from above, $E_{Y,QFT} = 1$. The exertion that previously guaranteed him an A now gives him a 50% chance of an A and a 50% chance of a B. To evaluate utility and find the new equilibrium, we should find formulae for the probabilities of each grade given effort. Applying the relevant constants and variables along with the cumulative density function[11] for our noise variable $\epsilon_{i,QFT}$, we get

$$P(A) = P\left(A_i^{\frac{1}{2}} E_{i,QFT}^{\frac{1}{2}} + \epsilon_{i,QFT} > \frac{1}{2}\right) \tag{2.16}$$

$$P(A) = P\left(\epsilon i, QFT > \frac{1}{2} - \frac{1}{2} E_{i,QFT}^{\frac{1}{2}}\right) \tag{2.17}$$

[10]The uniform distribution has the practical advantage that its cumulative distribution function can be expressed as elementary functions; that of the normal distribution cannot. The general character of our results will not be affected.

[11]

$$F(x) = \begin{cases} 0 \text{ if } x < \frac{1}{4} \\ 2x - \frac{1}{2} \text{ if } x \in [\frac{1}{4}, \frac{3}{4}] \\ 1 \text{ if } x > \frac{3}{4} \end{cases}$$

$$P(A) = \begin{cases} 0 \text{ if } (A_i^{\frac{1}{2}} E_{i,QFT}^{\frac{1}{2}}) < \frac{1}{4} \\ 2(A_i^{\frac{1}{2}} E_{i,QFT}^{\frac{1}{2}}) - \frac{1}{2} \text{ if } (A_i^{\frac{1}{2}} E_{i,QFT}^{\frac{1}{2}}) \in [\frac{1}{4}, \frac{3}{4}] \\ 1 \text{ if } (A_i^{\frac{1}{2}} E_{i,QFT}^{\frac{1}{2}}) > \frac{3}{4} \end{cases} \tag{2.18}$$

$$P(B) = 1 - P(A). \tag{2.19}$$

We'll use the standard expected utility calculation to determine what the new optimal course might be.

$$E(U) = P(A)U(A) + P(B)U(B) \tag{2.20}$$

Evaluating the possible cases, we find Yojimbo's new utility-maximizing effort is 0. We can think of this as a consequence of risk-aversion. With the introduction of noise, exerting effort becomes akin to buying a lottery ticket: certain cost for uncertain gain. It's no longer worthwhile.

We've been supposing Yojimbo is risk-averse, but suppose he's risk-neutral with a utility function like the following:

$$U_{Y,QFT} = G_{Y,QFT} - E_{Y,QFT} \tag{2.21}$$

Then a different picture emerges. Whereas without noise, Yojimbo might have exerted 0 effort, with noise he'll exert $\frac{1}{4}$. The larger lesson is that when noise is involved there are two effects at work, one which tends to increase effort and one which tends to decrease it, and which predominates is a somewhat unpredictable consequence of the degree of risk tolerance.

2.4.3 Are noisy grades desirable?

As we saw above, introducing noise into the performance function (equivalent to adding a random element to grading) can increase students' incentive to work above what they would normally exert in an imperfect grading system. This might be tempting, but we must remember that noise will sometimes reduce work below desirable levels, too, and that it's hard to predict when that will be the case.

Even if we could precisely identify the circumstances in which noise would produce more work, though, it probably wouldn't be a good tool for the purpose. The other consequences of noisy grading are just too bad. For one thing, random grading is inaccurate and seems unfair. Even if each student's *expected* grade is the grade they deserve, significant noise means that many students will not actually get the grade they deserve. As a result, some spots at competitive grad and pre-professional schools, and employers, will go to candidates who are not the most deserving while their more accomplished peers will have to settle for less, and we will have to reconcile ourselves to better students routinely losing out to worse in everyday contexts. The loss in accuracy (to use the language of our information content discussion) brings with it real losses in both equity and efficiency that grow as the magnitude of its incentive effects grow.

Probably most damning, though, is that the effect we might want noise to produce can be achieved without actually introducing randomness into the grading process. If noise improves incentive to work, it does it by creating uncertainty in students' minds about the grade they're likely to receive.

So if there were uncertainty in students' minds about the value of their performance function, that would produce the good effects of noise without introducing the sacrifices of efficiency and equity associated with randomness in assigning grades. We've left this out of our model for simplicity (and because mathematically it's similar to grading noise), but it's a feature of most real life classes; students often have only a general idea of how well they're doing in a class.

In conclusion, then, although the incentive properties of noise are interesting, it seems pretty clear that less noise is better.

2.5 Work and class selection

At 2.3 we showed students will choose to register for different classes based on their expected grade level, and above we showed that features of classes' grading regimes have significant effects on student behavior *within* each class. These two classes of observations deserve to be put together. In this section, we will investigate how grading regimes can influence students' choice between two classes even when their expected grades in those classes are the same. In particular, when students choose classes they'll consider not just the expected grade levels but how easy or difficult it is to achieve those levels. A clear example: Xenia is choosing between two classes, class 1 and class 2. Whichever class she takes, she'll get an A, but, because of differences in subject matter, in class 1 she'll get an A without working while in class 2 she'll need to work for it[12]. *Ceteris paribus*, Xenia will

[12]This could easily be the case if, for example, Xenia has high ability, and performance in class 1 is based disproportionately on ability while performance in class 2 is based

clearly choose to enroll in class 1. In this example, it would have been a feature of the performance functions of the two classes (representing facts about the subject matter) that made the difference in Xenia's decision, but features of how the class is graded could have similar effects; examples like Xenia's can be generated for most of the phenomena we've discussed above.

So if Willard is choosing between class 1 and 2 as given in 2.4.1, where class 1 uses a grading system with 2 subintervals and class 2 uses a grading system with 3 subintervals, and Willard would get the same grade in either class but have to work harder in class 2, he'd choose class 1. Or if Yojimbo is choosing between two classes that are identical except that one exhibits performance noise, and Yojimbo is risk-averse, he'll choose the noise-free class even if the expected grade is the same in order to avoid the chance of a utility-cramping loss.

The point of all this is that grading system imperfections have impacts that reach beyond the grades for any individual class by affecting course selection decisions. This makes the problems they cause both more serious and less tractable than they otherwise might seem. They are more serious because (for example) "easy" grading in just a few classes, or a single department, can affect the educational mission of an entire institution by drawing students out of the classes they would be best served by taking with the lure of better grades or less work – even though the root of the problem is not in course design, or material, or quality of instruction, but just in grading regime. They are nearly intractable because improving the grading system of any particular class or classes only deepens the systemic problem,

disproportionately on effort.

that disparities among grading systems distort course selection. Further, individual professors may be discouraged from improving their own grading systems for fear that students will leave their classes for easier ones sporting worse grading systems[13].

2.6 Robustness

Having developed the consequences of our model, it makes sense to ask which of the effects we've observed are real and which are model-dependent. The arguments about information content are nearly tautological: all other things equal, level of detail (precision) trades off against the degree of confidence in that detail (accuracy). The basic theory of course selection outlined is also nearly unassailable: all other things equal, students choose the courses that earn them higher grades for less work. To be sure, the assumption – "[a]ll other things equal" – rules out a lot of factors that really are relevant to course selection, like signaling value to transcript readers, interest in the material, teacher quality, and so on, but the mere fact that these are real considerations that in real life sometimes override the decisions laziness and grade-greed would dictate does not mean that our analysis is incorrect[14].

[13] Research suggests this is just one of several incentives that keep professors from grading their classes more strictly (and thereby better). [refer to that paper about teaching evaluations, etc.]

[14] If these factors could be shown to actually *wipe out* considerations like expected grade and expected work, then the analysis given above might be wrong, but they can't. For example, it's true that transcript readers can sometimes guess which classes were the hardest and weight those grades more heavily accordingly, which would tend to counteract students' tendency to choose the easiest courses. But transcript readers don't have that much information about individual courses (we ourselves have access to thousands of data points but still can't definitively rank classes by difficulty), so students still respond to differences in difficulty.

Similarly, even if we assume a radically different incentive structure for students' effort, our main results about the incentive to work should still hold (e.g., even if we assume a utility function that's increasing on effort independent of grade up to a certain amount of effort, so that students would do some work just because they enjoy it, our results clearly apply once students are over that minimal amount of work). And we've assumed relatively little about performance noise: it can be applied to any performance function, and, while we've assumed uniformity for convenience, any reasonable probability distribution should have similar effects. In sum, although the models help illustrate the way our results emerge from simple assumptions, the results are not dependent on a tight correspondence between model and reality.

2.7 Implications for policy goals

The arguments above lead us to two clear goals for a grading system:

Good effort Students should do enough work in their classes (based perhaps on the amount they would do under perfect grading).

Good selection Students should select classes without being influenced by grading system. They should choose based on personal preference, interest, skill and talent, the amount of work required to do a good job (not necessarily the same as the amount of work necessary to get a good grade), etc., not on how the class's performance interval is subdivided.

Chapter 3

Empirics

Now that we know what we want in courses' grading systems, we have a lens through which to view empirical data about students' grades. Below, we'll examine grade data from Amherst College and try to characterize it in the terms developed above. How fine-grained is each department's grading system? How much performance noise is there? Are these likely to result in problematic incentives?

3.1 Data sources

In examining grading patterns and their meaning for class difficulty, student incentives, and so on, I am fortunate to have access to a wealth of Amherst College grade and demographic data assembled by Harrison Gregg and the Office of Institutional Research, and further processed by Prof. Steve Rivkin. The records (which are anonymous) include every grade every student made in any class taken at Amherst during a period of several years, as well as some general information about each student (athletic participation, major, ethnicity, etc.); in total, the Amherst careers of several thousand students are covered in whole or in part. In order to meet ethical standards laid out by Dean of Faculty Greg Call, I will only use data on graduates (rather than current students). This is only a minor inconvenience, as the data covers several thousand recent graduates. I discuss a handful of individual departments using pseudonyms: "Sci" is a science or math, "Soc" is a social science, and "Hum" is in the humanities.

3.2 Average Grades

Some variation among departments is to be expected, but large differences along any of a number of dimensions could be meaningful and indicative of problems. For example, big differences in average grade between departments may be indications that it is easier to get higher grades in the higher-grading departments, causing students to enroll in classes in the higher-grade department more than they ought [Sabot and Wakeman-Linn1991]. A quick review of the Sci, Soc and Hum departments reveals a disparity (see table, p. 32). The numerical grade (row 1) ranges in value from 0 to 14, with 0 an F and 14 an A+. (I've assumed the scale is linear for purposes of computing the mean and later regressions.)

One obvious pattern in the average grade figures is that, if a department has high average grade, its average grade also has low variance (this relationship holds for any comparison among the Sci, Soc, and Hum departments). This is consistent with the observation that "grade inflation" is better de-

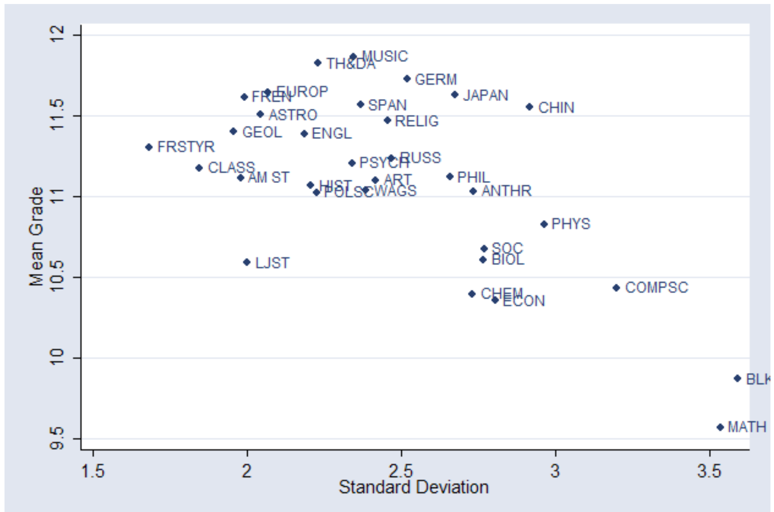

Figure 3.1: Departments scatterplotted by mean and standard deviation of grade.

scribed as "grade compression" [Ostrovsky and Schwarz2002, 2]. A simple model of grading makes it clear why that's the case: assume the teacher wants to give the best student in her class the best possible grade (an A+). Now, for a given distribution of student performance, the lower the grade the teacher gives the worst student, both the lower the mean grade in the class will be, and the higher the deviation about that mean. According to this view, then, departments that grade higher should necessarily have lower variance of grading. Figure 3.1 plots this relationship for all departments.

Of course, it's possible that inter-department grade differences are accounted for by corresponding differences in the quality of the students in those departments, even given the foregoing linkage between average grade and grade variance (in a class consisting wholly of high quality students, the teacher will likely want to give the worst student a relatively high grade).

If mean quality were higher in the higher-grading departments, then, those departments would grade higher even if their grade system were the same as lower-grading departments, and likewise if the departments with higher variance of grades also contain students with higher variance of quality. Fortunately we can test this observation by checking the following:

1. Do higher-grading departments have higher average quality?

2. Do departments with higher variance of grades have higher variance of quality?

3. Among students of the same quality, is the interdepartmental grading gap smaller? (If quality differences are a significant explanation for interdepartmental grading differences, that gap should be quite small).

We can check 1. and 2. by appeal to summary statistics; 3. we postpone to the section on regression analysis.

Table 3.1

	Sci	Soc	Hum	All depts[a]
Numerical grade	9.57 *(3.53)*	10.35 *(2.81)*	11.39 *(2.19)*	11.04 *(2.54)*
Mean of class sd	2.62	1.73	1.67	2.00
Sd of class means	1.39	1.15	.90	2.25
Student GPA	11.08 *(1.36)*	11.18 *(1.07)*	11.42 *(1.09)*	11.28 *(1.14)*
Mean of class sd	1.00	.96	.91	.95
Sd of class means	.73	.66	.52	.77
Math SAT	712.85 *(65.15)*	703.81 *(57.23)*	691.10 *(61.35)*	695.34 *(63.49)*
Mean of class sd	95.24	70.78	66.72	70.34
Sd of class means	61.17	51.64	31.87	73.47
Verbal SAT	701.83 *(72.36)*	687.43 *(70.53)*	718.79 *(62.90)*	703.78 *(67.04)*
Mean of class sd	98.94	88.89	71.71	73.94
Sd of class means	51.35	65.69	33.27	72.76

Source: Amherst College data.

[a]This figure is computed by considering all the students (or all the classes) in all departments, rather than by two-stage averages (e.g., averaging first within, then across departments).

The main difficulty here is that there are no really good external measures of student quality, no perfectly reliable method for determining what grade a student *should* get in a particular class (if there were, would we need actual grades?). A few approaches are possible, however:

Cumulative average grade (GPA). If we assume the component of student quality that applies in every class the student takes (general intelligence, study skills, etc.) is large (and if all classes are graded on similar scales), then we'd expect GPA to be a pretty strong predictor of grade in any given class, and a good measure of student quality. The difficulty is that the GPAs of students who take many courses in high-grading departments will be higher by that mere fact, introducing a bias into the estimate of quality that could make it seem that such departments' grades are justified by high student ability. In the language of our theoretical model, GPA might be thought of as a measure of ability *and* the tendency to exert effort (even though we can't measure effort in each class).

SAT scores. Studies find SAT scores to be excellent predictors of grade performance, which suggests that they are a good measure of student quality (think of them as estimators of a "natural ability" component). In the language of our theoretical model, SAT might be thought of as a measure of ability, A_i.

Admissions reader rating. When the Admissions Office at Amherst College reads applicants' files, it assigns each a numerical "reader rating" based on the expected academic strength of the applicant ([Symonds2006]). The number ranges from a best rating of 1 to a worst of 7 (with some variation from year to year). No 6s or 7s are admitted. The reader rating is tied closely to SAT, but it's not wholly determined by it, and may serve as a better predictor of performance, since admissions readers also incorporate information from high school transcripts, essays, and so on. Like GPA, reader rating is a way to try and get at $A_i+E_{i,j}$.

Our table includes average GPA and verbal and math SAT, by department and for the whole school. These averages are weighted by class rather than by student; if a student took two math classes, his credentials count twice as much toward the average math department credentials as a student who took only one. There are no clear trends in credentials. Math SAT is highest in the Sci department and verbal in the Hum department (a selection effect), but the difference is small, and total SAT is nearly the same, regardless (it's somewhat lower in Soc, but Soc grades higher than Sci, anyway). If the SAT is our measure of ability (whatever that means), then, ability differences don't appear to explain the grade difference among the departments we're examining. This doesn't, however, eliminate the possibility that a different sort of difference in the characteristics of students could explain the disparity; what if Hum students have better study habits, for example, a difference likely not captured by SAT score?

Better study habits, or any other population-level difference in skills applicable in many academic subjects (not just a particular department) *would* be captured by students' GPAs, so we can examine those figures to give a fuller answer to the problem posed above. Inspecting the GPA figures in the table, there does appear to be a slight trend in GPA (with Hum students having the highest and Sci students the lowest), but the differences are small, and dwarfed by the interdepartmental grading differences. It seems implausible that the differences in quality indicated by such small differences in GPA could account for the big gap in average grade, although without a regression analysis it's hard to answer that question definitively[1].

[1]As previously noted, there's what one might call an endogeneity problem here: GPAs

Even though differences in average quality don't appear to be the answer, though, differences in the variance of quality might be. However, the data don't appear to support such a conclusion.

3.3 The Course Structure of Departments

There's another possible explanation for the higher variance of grades in some departments that's completely independent of the ones offered above: different classes in the same department could be graded differently. We'll examine how this might explain the observed numbers, what would need to be true in order for it to be intuitively justified, and what our summary statistics say about the reality.

We've been effectively assuming high variance of grades for a department means that a wide range of grades is used in all classes in that department, but that need not be. Consider the limiting case in which every student in each class in a department gets the same grade (for example, the course number). If there are many different courses in the department (with widely varying numbers) relative to the total number of students, grades in the department will exhibit a high variance even though there's zero variance in every class. Fortunately for our analysis above, this doesn't appear to be the case. As figure 3.2 shows, when we consider each class individually, there's still a wide range in the standard deviation of grade, and the relation between

are influenced by the grading standards of departments in which students take classes, so a high-grading department could inflate its students' GPAs in a way that makes it seem like they are better qualified than they are and appears to justify the department's high grades. However, since students who take easy classes in a given department will take easier classes overall for the same reasons, the problem is hard to eliminate. Instead, we can appeal to reader rating.

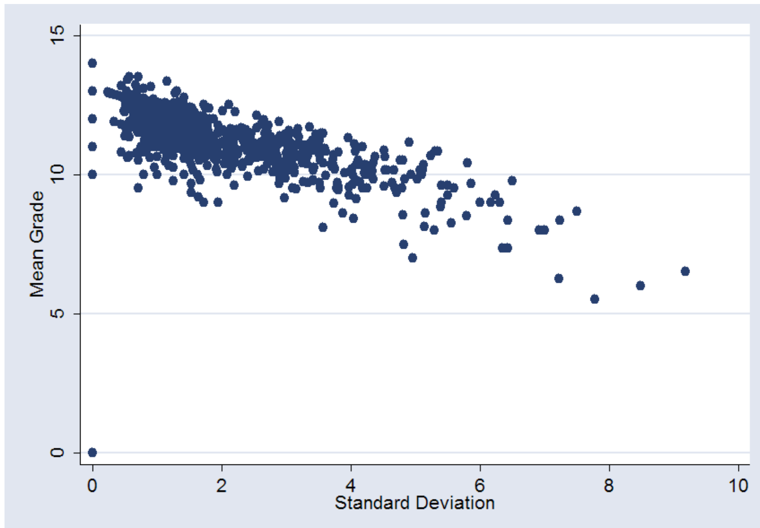

Figure 3.2: All courses scatterplotted by mean and standard deviation of grade.

mean grade and standard deviation of grade we observed among departments still holds. Furthermore, in the table on p. 32, we produce values for mean and deviation both within and across courses for the departments covered, and it's clear from these that variation among classes cannot tell the whole story. In fact, the pattern outlined above holds for variation both within and across classes: within Sci classes, there's likely to be more variation (of grades and of quality measures) than in Soc, in whose classes there's likely to be more variation than in Hum. Likewise, there is more variation among the class means of classes in Sci than in Soc and in Soc than in Hum.

This gives us an interesting and non-obvious point about the course structure of departments: the departments whose classes employ broader grade scales also have more stratified course structures. The difference – both in grading and in student quality – between Sci 5 and Sci 42 is much greater than between most Hum classes, for example. It's hard to see what the connection is, but we can hypothesize: student quality will be more stratified by class when students have the incentive to sort themselves into classes by ability, and that incentive will be much stronger in departments that grade more harshly, since the prospect of sorting into a too-hard class is much less attractive if it is likely to result in a larger grade penalty.

3.4 Regressions

In this section, we'll estimate a few simple models wherein numerical grade is expressed as a function of quality measures. The measures used, as discussed above, are SAT, GPA, and dummies for reader rating, but of the three, reader rating is the best, because it can incorporate some reasonable expectations about effort (unlike the SAT) without the course-selection endogeneity problems of GPA. The results are roughly the same regardless of which measure is used, so I will only report reader rating regressions below. There is, however, one inconvenient fact about reader rating: what a particular reader rating means changes from year to year, and there have been a handful of major regime changes, as well. We can't make the restriction that a given reader rating means the same thing every year.

3.4.1 Grading system

The population trends in grade tend to support the story we've been telling, that interdepartmental differences in grade mean and standard deviation are due at least in part to differences in grading system (rather than, say, differences in quality), but we haven't truly ruled out quality as an explanation. Above, I argued that we could test this by asking whether, among students of the same quality, interdepartmental grade gaps are smaller? The obvious way to answer this question is to regress grade on quality measures with dummies for department; we'll use our old favorites, Sci, Soc, and Hum, again. Using reader rating as our measure of quality, the coefficients to the dummies for Soc and Hum are both large and significant at the 1% level, meaning that ability differences probably cannot account for the grade differences among the departments.

Table 3.2

	Coef.	t stat.
Soc	.84	7.35
Hum	1.69	15.30

Source: Amherst College data.

Combining these findings with everything else we know, then, it seems safe to conclude that grade differences among departments (and with more certainty among these departments) are accounted for at least in part by differences in grading system, in the sense we discussed in the theory section. Some departments have finer, and some less fine grading systems.

3.4.2 Signal and noise

In our theoretical discussion, we identified another dimension along which classes could differ, separate from grading system: magnitude of performance noise. In our theoretical model, performance noise is what accounts for differences in grade after ability and effort are accounted for. If classes and departments vary in the magnitude of their performance noise, that would influence course selection in adverse ways (etc.). The obvious way to test the magnitude of performance noise is to use regression models to find out how much of the variation in grade is accounted for by ability and effort (i.e., quality) – what's left over is noise. To compare the magnitude of performance noise across departments, then, we can compare goodness of fit of a model that estimates numerical grade as a function of ability.

Table 3.3

	(1)	(2)	(3)	(4)
Sci	.16	.44	.03	.88
Soc	.13	.34	.05	.75
Hum	.07	.24	.08	.67

Source: Amherst College Data.

The table above gives R^2 values for the regression of numerical grade against reader rating (column 1). The results indicate that quality seems to explain the largest share of grade variation in Sci, and the smallest in Hum, which suggests that Sci has the least noise and Hum the most. Of course, it would almost certainly be wrong to conclude that the share of grade due to actual performance noise in these departments is $1 - R^2$. If that were the case, noise would account for 84% of grades even in the least noisy department. Part of the problem is that reader rating doesn't do a great job of capturing student effort. More generally, though, the issue is that there are some factors (probably not noise) that account for a lot of grade variation but that our model doesn't capture, which limits the validity of our results. One way to make the results more robust is to fit a more comprehensive model; I've included the results for the regression of grade on reader rating, SAT and GPA above (column 2)[2].

The nagging problem of course structure remains, however: it could be that the appearance that the Sci department pays higher dividends to ability is an illusion if what's actually going on is simply that higher-level courses are graded more easily, and higher-quality students sort into those courses. We can try to rule out this and other course structure-based explanations by estimating a model with course fixed effects and finding out how much of the variation is explained by course. Column 3 above gives the proportion of variation explained by course. If course structure was confounding our attempts to measure noise, we would find that course had the most influence on grade in Sci and the least in Hum, but in fact we find the opposite.

Could what looks like noise in Hum department grades actually be the influence of unobserved differences in student quality, some kind of department-specific hidden ability? In order to account for this possibility, we can use average grade in department as an instrument for department-specific hidden ability and see whether that changes the interdepartmental contrasts we've observed. Column 4 above reports the R^2 values for a reader rating model that includes in-department average as an explanatory variable. Sci still appears to be the least noisy department, and Hum the most. What all of this means is that our preliminary conclusions about noise were probably correct: there's significant interdepartmental variation in performance noise, with Sci exhibiting less of it and Hum more.

This points to a curious contrast. Recall: among Sci, Soc, and Hum, Sci exhibited the greatest variance of grade and Hum the least. But at

[2]SAT models exclude students who took the ACT instead.

the same time Sci has the least (and Hum the most) *bad* variation. It's not obvious why the departments with the finest grading systems should also be the least noisy and those with the coarsest also the noisiest. One theory: professors are in competition with each other to attract students to their classes, and high grades or low noise can each attract students. Professors in high-grading departments can afford to grade more noisily, because even if it drives a few students away, plenty will stick around for the high grades. What this (admittedly speculative) hypothesis means if true is that interdepartmental grading and noise differences are in a self-reinforcing equilibrium: high grades are linked to high noise which is linked to high grades. One potential upside is that it might be sufficient to fix just one of the problems: if all departments effectively used the same grading system, no departments would be able to afford the enrollment cost of noisy grading.

Chapter 4

Conclusions

When we apply our new theoretical tools to our new factual knowledge, the results are somewhat alarming, or at least disappointing. Differences in grading system and performance noise both have somewhat unpredictable but clearly deleterious effects, and the Amherst College curriculum seems to have both in abundance. It would be easy to throw our hands up in despair since of course it would be a grievous violation of faculty autonomy to actually reach into classes and manipulate their grading systems, but I think we're obliged to at least review the available solutions before giving up. As I alluded to earlier, several papers have been published suggesting that interdepartmental grade inconsistencies can be partly resolved by employing some form of grade post-processing: adjusting grades or GPAs by subjecting them to weighting mechanisms after they've been assigned. What follows is a brief review of that literature, with directions for further work.

4.1 What solutions are available?

There are a wealth of similar solutions, and it's hard to distinguish among them. Here are some highlights: the simplest solution is probably the one offered by [Felton and Koper2004], which the authors call a "real GPA" – each student receives a real grade equal to the ratio of his nominal grade to the average nominal grade for the class, normalized so that the mean student receives a C. This method means that every class has a grade distribution with certain well-defined characteristics, but it doesn't solve all the problems because it doesn't account for differences of difficulty among classes. Another simple but more robust solution is presented in [Caulkin and Wei1996]. Caulkin et al. use a handful of different linear models based around the assumption that a student's grade in every class is a combination of a (one-dimensional) achievement or quality factor, a class-specific factor, and an error. [Johnson1997] is a refinement of the same concept that uses a more complicated, Bayesian estimation model.

4.2 How do we evaluate solutions?

Most papers make an attempt to evaluate their proposed solutions by dead reckoning and economic reasoning. However, the only quantitatively serious evaluations seem to rely testing the fit of regressions of adjusted grade numbers on variables expected to predict performance (i.e., quality measures). Both [Caulkin and Wei1996] and [Johnson1997] test how well high school GPAs and SATs predict their adjusted (college) grades, and compare the results to those of the same regressions using unadjusted grades as the dependent variable. This is a pretty good methodology for evaluating how well a particular grade adjustment aligns with student quality. However, we may be able to do better. With the theoretical framework we've developed, we should be able to give rigorous answers to questions about what effect a particular grade adjustment will have on student incentives – ultimately one of the most important considerations when it comes to which adjustment to adopt.

Now that we know what we do about the theory and reality of grading, we are better equipped to evaluate proposed solutions to grading problems, or develop new ones. It could make a tremendous difference.

Bibliography

[Caulkin and Wei1996] Caulkin, J. Larkey, P., and J. Wei. 1996. Adjusting GPA to reflect course difficulty. Working paper, Heinz School of Public Policy and Management, Carnegie Mellon Univ.

[Dubey and Geanakoplos2004] Dubey, Pradeep, and John Geanakoplos. 2004. Grading Exams: 100, 99, ..., 1 or A, B, C? Incentives in Games of Status. Cowles Foundation Discussion Papers, vol. 1467.

[Ellenberg2002] Ellenberg, Jordan. 2002. Don't Worry About Grade Inflation. Slate.

[Felton and Koper2004] Felton, James, and Peter T. Koper. 2004. Real GPA and Nominal GPA: A Simple Adjustment that Compensates for Grade Inflation.

[Johnson1997] Johnson, Valen E. 1997. An Alternative to Traditional GPA for Evaluating Student Performance. Statistical Science 12 (4): 251–278.

[Kelley1975] Kelley, Allen C. 1975. The Student as a Utility Maximizer. Journal of Economic Education 6 (2): 82–92. FLA 00220485 Joint Council on Economic Education Copyright 1975 Heldref Publications.

[Ostrovsky and Schwarz2002] Ostrovsky, Michael, and Michael Schwarz. 2002. Equilibrium Information Disclosure: Grade Inflation and Unraveling. Working Paper, Harvard University.

[Sabot and Wakeman-Linn1991] Sabot, R., and J. Wakeman-Linn. 1991. Grade Inflation and Course Choice. Journal of Economic Perspectives 5 (1): 159–170.

[Spence1973] Spence, Michael. 1973. Job Market Signaling. The Quarterly Journal of Economics 87 (3): 355–374. FLA 00335533 Harvard University Press EN Copyright 1973 The MIT Press.

[Symonds2006] Symonds, William. 2006. Campus Revolutionary. Business Week, no. 9.