View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Research Papers in Economics

GRADING EXAMS: 100, 99, ..., 1 OR A, B, C?
INCENTIVES IN GAMES OF STATUS

By

Pradeep Dubey and John Geanakoplos

July 2004

COWLES FOUNDATION DISCUSSION PAPER NO. 1467

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY

Box 208281

New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Grading Exams: 100, 99, ..., 1 or A, B, C? Incentives in Games of Status

Pradeep Dubey and John Geanakoplos

June 18, 2004

Abstract

We show that if students care primarily about their status (relative rank) in class, they are best motivated to work *not* by revealing their exact numerical exam scores $(100, 99, ..., 1)$, but instead by clumping them in broad categories (A, B, C). If their abilities are disparate, the optimal grading scheme awards fewer A's than there are alpha-quality students, creating small elites. If their abilities are common knowledge, then it is better to grade them on an absolute scale (100 to 90 is an A, etc.) rather than on a curve (top 15% is an A, etc.). We develop criteria for optimal grading schemes in terms of the stochastic dominance between student performances.

Keywords: Status, Incentives, Education, Grading, Wages

JEL Classification: C70, I20, I30, I33

1 Introduction

Examiners typically record scores on a precise scale 100, 99, ..., 1. Yet when they report final grades, many of them nowadays tend to clump students together in broad categories A, B, C, discarding information that is at hand. Why?

Many explanations come to mind. Less precision in grading may reflect the noisiness of performance: a 95 may be statistically insignificantly better than a 94. Alternatively, the professor may require less effort in dividing students among three categories rather than a hundred. Finally, it may be that lenient grading is a device by which professors lure students into their class; unable to call an exam with 70% correct answers a 95 instead of a 70, they call it an A instead.

We call attention to a different explanation. Suppose that the professor judges each student's performance exactly, though the performance itself may depend on random factors, in addition to ability and effort. Suppose also that the professor is motivated solely by the desire to induce his students to work hard. Third, and most importantly, suppose that the students care about their relative rank in the class, that is about their *status*. We show that, in this scenario, coarse grading motivates status-conscious students to work harder.

Status is a great motivator, though it is largely absent from economic models. For students competing for scarce positions in the job market or in higher education, status or rank is often the decisive factor, and hence it naturally becomes the student's major concern. For workers, such as professors themselves, honors conferring status but little remuneration often bring forth the greatest effort. Ranks and titles are ubiquitous, in academia, in the armed forces, in corporations, and in public bureaucracies. They define a hierarchy which, even when its original purpose might have been organizational (say to signal lines of authority), always creates incentives for people to exert effort in order to obtain higher status.

One might think that finer hierarchies generate more incentives. But this is often not the case. Coarse hierarchies can paradoxically create more competition for status, and thus provide better incentives for work, under certain conditions.

The advantage of coarse grading can most succinctly be seen with two students who have *disparate* abilities, so that the strong student achieves a random but uniformly higher score even when he works little and the weak works hard. With such disparate abilities and fine grading, the strong will come ahead of the weak, regardless of their effort levels. Since they care only about rank, *both* will shirk. But we show in Section 3 that appropriately-defined coarse grades can inspire the weak to work, for then he stands a chance to acquire the same status B as the strong. This in turn generates the competition which spurs the strong to also work so that, with luck, he can get an A and distinguish himself from the weak. Notice that very coarse grading would not do the job since then nobody has anything to gain by improving his score. Optimal grading must be coarse, but not too coarse.

It should be emphasized that coarse grading does not involve what are commonly called handicaps. It is of the essence of handicaps that they discriminate between contestants by bestowing an advantage on the weak. Handicaps thus presume knowledge of the contestants' abilities, as well as the "legality" of the discrimination. The grading we describe in this paper is, in contrast, completely anonymous in that grades depend only on the exam scores of the students and not on their names. It is also monotonic in the scores: if a student gets a better score than another, he is awarded at least as good a grade. On either count, handicaps are ruled out since they would necessarily entail an artificial boost to the score/grade of the weak student.

There can be schemes, other than coarse grading, which also generate competition between the students without recourse to discriminatory handicaps. For instance, the professor might add anonymous noise to each student's score with enough variance so that, if the weak student works, he stands a reasonable chance to equal or overtake the strong. While theoretically intriguing, this scheme is so far removed from reality — and, some might argue, from even being made acceptable as reality — that we do not pursue it (or other like-minded schemes) here.[1] Our focus is on the familiar and time-honored tradition of assigning letter grades to the numerical exam scores, with no tampering of the scores.

Though we do not impose extraneous noise on students' performance, it is critical for our grading scheme that there be *some* degree of intrinsic randomness (albeit arbitrarily small) in the performance. This hypothesis, in our opinion, should strain

[1] There *are* situations when noise can be "natural." In Dubey–Wu (2001) and Dubey–Haimanko (2003), noise was introduced by varying the sample size on the stream of outputs produced by agents.

no one's credulity. In the rare situation when performance is a deterministic function of effort, any grading coarse enough to incentivize the weak by enabling him to be pooled with the strong, will destroy the possibility for the strong to separate himself from the weak, leaving the strong with no incentive to work. To restore the scheme, it is needful to add a *touch* of randomness into the scores from the outside. The *conjunction* of a little randomness and appropriately coarse grading makes for simultaneous incentives for all the students.

Many standard games, played in the parlor or the sports field, fit our general framework. Performance can often be very finely calibrated (sometimes literally by "the score"). But this is overlooked and the outcome is just declared to be a win, draw or loss. The frequent possibility of draws is akin to a coarsening and is surely introduced to sustain competition. In parlor games, randomness is often reinforced by inserting moves of chance (throw of the dice, shuffling of the deck, etc.) to further bolster the possibility of victory for the weak when he is overmatched. These twin principles of coarsening and randomness have influenced the design of games since antiquity. We bring them to bear on the game of status that is induced among students taking an exam. And we call attention to the universal prevalence of games of status in every day life.

Our main theme, as was said, is that the optimal design of exams often leads to a coarse partition of numerical scores into letter grades. This no doubt reduces the screening content delivered by schools. But our analysis reveals that if the schools sought to convey more information about the quality of their students, they would produce students of lower quality!

In Section 3 we characterize the optimal absolute grading scheme for an arbitrary number of students K and N of two disparate abilities, and then we show how to extend the analysis to three or more ability types. Throughout most of the paper we assume that the distribution of ability types and the distribution of scores each type will get on the exam, conditional on effort levels, is common knowledge among the students and the professor. By virtue of repeated meetings of the class, perhaps even over many years, it is not unreasonable to suppose that these distributions can be fairly well estimated by the professor and the students alike, so that this assumption is not so untenable. Each student also knows his own ability type. Our first and most important conclusion is that in order to create the largest incentives to work, the professor should use coarse grading.

Our second conclusion is that in apportioning status to foster incentives to work, the optimal grading creates small elites, excluding many from membership who have equal abilities and have also worked hard. In a population made up of equal numbers of students of disparate abilities, say alpha and beta and gamma, fewer A grades will be given than B's, and fewer B's will be given than C's. In particular, though they all work hard, only some alphas get A and only some betas get B. Furthermore, we show that as the proportion of alpha ability students increases, the fraction of them who should be given A's decreases. No matter what the composition of student ability types, no more than 50% of them should be given A's. If less able students have higher costs from studying hard, as Spence (1974) suggested, then the pyramiding becomes still more extreme.

The hypothesis of disparate abilities is strong, but not as strong as it seems, and can be plausibly interpreted. For example, one might imagine that students have many effort levels, and that when the alpha students exert their second best effort they will come ahead of the beta students, no matter how hard the betas work or how lucky they get. If the professor wants to motivate each student to do his *very* best, then our previous analysis holds without change. But we consider two other situations as well.

Coarse grading can provide better incentives to work even when students are *ex ante* identical. In Section 4 we provide criteria for an optimal grading partition in this setting. The key analytical concepts in this analysis are stochastic dominance and uniform stochastic dominance.[2] We show that a partition is optimal if inside each cell, the shirker's performance uniformly stochastically dominates the worker's, while across cells the worker's uniformly dominates the shirker's. Fine partitions are typically not optimal. For example, if the performance of a hard worker has normal distribution $N(\mu, \sigma)$ with mean μ and standard deviation σ, while the performance of a shirker is given by $N(\hat{\mu}, \hat{\sigma})$, then the optimal grading scheme will make perfectly fine distinctions on one tail, and mask all scores on the other tail into the same grade, provided that $\sigma \neq \hat{\sigma}$. (If $\sigma = \hat{\sigma}$ and $\mu > \hat{\mu}$, then perfectly fine grading is optimal.)

In Section 5 we consider the mixed case, where students are neither identical nor disparate, having significant overlap in their performance. We imagine an exam with K questions, and students $n \in N$ who have probabilities $p_1 < p_2 < \cdots < p_N$ ranging uniformly on $[0, 1]$, of answering each question (independently) correctly, if they work hard. If a student n shirks, his p_n is reduced to p_{n-1}. It turns out that coarse grading increases the incentive for the best and the worst students to work hard, though diminishing the incentives of students in the middle. But since the middle students already have a big incentive to work hard, the optimal grading partition to incentivize *every* student to work hard has to be coarse.

In moving from scores to grades, professors can grade absolutely (say 90 to 100 is an A) or "on a curve" (say the top 15% get an A). Given that the students only care about their relative rank, which kind of grading is better? We show in Section 6 that if the students have identical abilities and utilities, then absolute grading is always better than grading on a curve. This principle may be valid with heterogeneous students, but we leave its exploration for future research. It breaks down, however, when the professor is significantly uncertain about the distribution of students' abilities (though still the professor must, by and large, devise coarse grades).

In Section 7 we consider the problem the professor faces to motivate his students to study for two exams, say when he gives a midterm and final. The problem is that if a student does very well or very badly on the midterm, he (or his rivals) will feel less incentive to work for the final if he or they are unable to affect their course grade. The only way around this problem is to weight the final more than the midterm, and

[2] Stochastic dominance frequently plays a role in economic modeling, but uniform stochastic dominance has been less often used.

to grade even more coarsely on the midterm than on the final, averaging the letter grades and not the scores from the two exams to obtain the final course grade.

Lastly, in Section 8 we observe that status pertains to situations far more general than assigning course grades. We show that an employer who can induce his workers to regard superior performance as an indication of status will be able to get them to work harder and pay them less money. To minimize his cost, he should confine himself to a few grades of wages, even if productivity can be much more finely measured.

Society as a whole has much to gain by associating status with productivity. This insight was present and put into practice in the ancient world. As far back as 500 B.C., the Greeks were masters of designing competitions — for the best playwrights, the best artists, the best athletes and so on. Around the same time, at the other end of the civilized world of the day, contests of all kinds were also common practice in India, and included prominently among these were "Shastrarthas" or philosophical debates, with much prestige accorded to the winner. The ancients had clearly discovered that games of status often call forth the most heroic effort. (This is not to deny, of course, that calm Buddhistic detachment may also lead to inspired activity.)

This paper is a first step in analyzing games of status. We have pretty much characterized optimal grading partitions for the case of disparate students, and for the case of iid homogeneous students. We leave to future work the general case of heterogeneous (and overlapping) performances, and also the case where exam performances are correlated (for example, because everyone might find the exam unexpectedly hard).

2 Games of Status

In this section we precisely define what we mean by games of status, and the freedom the principal has to design them.

Imagine a group of students N who are taking an exam. Depending on their effort levels $(e_n)_{n\in N}$, they will get exam scores, $(x_n)_{n\in N}$, which might also depend on random events, such as whether they were lucky enough to have studied the material precisely relevant to the questions, or how they felt that day, or how accurately the professor corrected the exams. It is natural to assume that a student's score does not depend on others' efforts, but actually none of our mathematics requires it. When the score x_n depends (perhaps negatively) on the effort e_m, $m \neq n$, we can reinterpret our model as a parlor game. Given the exam scores $x = (x_n)_{n\in N}$, the professor must assign grades $\gamma(x)$. Students are assumed to care only about status (and not about the education they are getting). We capture this by assuming[3] that they obtain 1 utile for each student whose *grade* is strictly lower, and they lose 1 utile for each student whose grade is strictly higher.

We suppose that the students are told in advance how the professor converts scores to grades, i.e., they know γ. Absolute grading is achieved by specifying intervals of

[3] This is to keep matters simple. One could postulate β_{ij} utiles to i from beating j and γ_{ij} disutiles from being beaten by j, without changing the tenor of our analysis. Even aditively separable utilities are not essential.

scores corresponding to each grade, say $[90, 100]$ gives A, $[75, 90)$ gives B, and so on. Grading on a curve is based in contrast on relative performance alone, for example, that the top 10% of students get A, the next 20% get B's, and so on. Absolute and relative grading are quite different, though both are widely used.

What grading system γ *should* a professor use, if he wants to incentivize (whenever feasible) *all* his students to put in maximal effort? No matter what system he chooses, and no matter what efforts the students put in, total utility awarded via grades will be zero, since for every utile gained by a higher-ranked student, there is a utile lost by a lower ranked student. Indeed when students all work hard, their total net utility is minimized (since work inflicts disutility). Status seeking is the ultimate rat race!

Nevertheless, by the right choice of γ, the professor can often motivate his status-conscious students into working hard, and thus willy-nilly becoming educated.

2.1 The Performance Map

Let N be a finite set of *students* (players). The *strategy set* $E_n \subset \mathbb{R}$ of each student $n \in N$ consists of a finite set of effort levels that are w.l.o.g. identified with the disutility they inflict on n. Efforts lead to (random) performance scores. For $x \in \mathbb{R}^N$, the nth-component x_n of x represents the score (output) obtained by n. Let $E \equiv \mathsf{X}_{n \in N} E_n$ and let $\Delta(\mathbb{R}^N)$ be the set of probability distributions on $\mathbb{R}^N$. The *performance map*

$$\pi : E \to \Delta(\mathbb{R}^N)$$

associates stochastic scores with effort levels. Here $\pi(e)$ gives the probability distribution of score vectors when the students put in effort $e \in E$.[4] We allow for the possibility that n's effort e_n might affect the score x_m of other students $m \neq n$.

2.2 Grading

Let $\mathcal{R}$ denote all possible orderings of N with ties allowed. There is a *grading map*

$$\gamma : \mathbb{R}^N \to \mathcal{R}$$

which weakly ranks students according to $\gamma(x)$ when the scores obtained are $x \in \mathbb{R}^N$. Each rank corresponds to a grade, and so coarse grading entails several ties for the same rank. We consider, in principle, only maps γ that are anonymous and monotonic: the grades depend on the scores, not on the names, and a higher score implies at least as high a grade. Our focus will be on two particular ways of generating γ.

2.2.1 Absolute Grading

Let $\mathcal{P}$ be a partition of $\mathbb{R}$ into consecutive intervals, each of which has nonempty interior and some of which are designated "perfectly fine." When an interval[5] $[a, b)$

[4] In the natural case (see our examples), higher effort levels tend to improve scores in the sense of first-order stochastic dominance.

[5] We use $[a, b)$ as a proxy for $[a, b)$, $(a, b]$, (a, b) or $[a, b]$. Our analysis works equally in all cases.

is designated perfectly fine, it is taken to represent the partition $\{\{x\} : x \in [a, b)\}$ consisting of singleton *cells*. An interval $[a, b)$ not so designated will signify the standard unbroken interval, and will also be called a *cell* in the partition $\mathcal{P}$. Notice that the cells of $\mathcal{P}$ form a totally ordered set.

Fix a partition $\mathcal{P}$ as above. Then for any $a, b \in \mathbb{R}$ we define $a \succ_{\mathcal{P}} b$ iff the cell in $\mathcal{P}$ containing a lies strictly above the cell in $\mathcal{P}$ containing b. This leads to an absolute grading $\gamma_{\mathcal{P}} : \mathbb{R}^N \to \mathcal{R}$ where $i \succ_{\gamma_{\mathcal{P}}(x)} j$ iff $x_i \succ_{\mathcal{P}} x_j$. Thus $\gamma_{\mathcal{P}}(x)$ coarsens the information in x, creating ties between players whose scores lie in the same cell of $\mathcal{P}$.

2.2.2 Grading on a Curve

Given scores $x = (x_n)_{n \in N} \in \mathbb{R}^N$, define the class rank $\rho_n(x)$ of each student n by

$$\rho_n(x) = \#\{j \in N : x_j > x_n\} + 1.$$

Several students may have the same class rank. We define the "grading curve" Q to be a consecutive partition of class ranks $\{1, 2, ..., |N|\}$. For any $x \in \mathbb{R}^N$, let $\gamma_Q(x)$ be given by

$$n >_{\gamma_Q(x)} j \text{ iff } \rho_n(x) <_Q \rho_j(x).$$

This defines the grading map $\gamma_Q : \mathbb{R}^N \to \mathcal{R}$.

In words, a grading curve is defined by the number n_A of students getting A, the number n_B getting B, and so on. The grades are obtained by ranking student exam scores, and taking the top n_A scores and giving all the students who got them A. If $k > n_A$ students tie with the top score, then all must get A, and the number of B's is diminished by the excess A's, and so on.

2.3 Utilities

The (exam) performance payoff to a student n from being ranked according to $R \in \mathcal{R}$ is

$$\#\{j \in N : n >_R j\} - \#\{j \in N : j >_R n\}$$

reflecting the fact that n gets a utile for each student he beats, and loses a utile for each student who beats him. He cares about status.

Thus a student n who exerts effort $e_n \in E_n$ and faces ranking $R \in \mathcal{R}$ gets net utility:

$$\#\{j \in N : n >_R j\} - \#\{j \in N : j >_R n\} - e_n$$

Notice again that the student is indifferent to learning. Had he put value on it, our task of incentivizing him to work would have been much simpler.

2.4 The Game Γ_γ

Fix a grading function $\gamma : \mathbb{R}^N \to \mathcal{R}$. Then, given effort levels $e \equiv (e_k)_{k \in N} \in E$, the *payoff* to $n \in N$ is his expected net utility:

$$\text{Exp}_{\pi(e)}[\#\{j \in N : n >_{\gamma(x)} j\} - \#\{j \in N : j >_{\gamma(x)} n\}] - e_n \equiv u^n_\gamma(e) - e_n.$$

Here $\text{Exp}_{\pi(e)}$ denotes expectation w.r.t. the distribution $\pi(e)$ over scores $x \in \mathbb{R}^N$ and $u^n_\gamma(e)$ denotes the *expected exam payoff* to n.

For $e \equiv (e_k)_{k\in N} \in E$ and $n \in N$, denote $e_{-n} \equiv (e_k)_{k\in N\setminus\{n\}}$. Recall that e is a *Nash equilibrium* (NE) of Γ_γ if the payoff each student n gets under e is at least as good as the payoff under (e'_n, e_{-n}) for all $e'_n \in E_n$.

Let $\tilde{e} \equiv (\tilde{e}_n)_{n\in N}$ be the strategy profile of *maximal effort*:

$$\tilde{e}_n = \max\{e'_n : e'_n \in E_n\}.$$

The key concern of our analysis is to design γ so as to ensure that $\tilde{e}$ is an NE — hopefully the unique NE — of the game Γ_γ.

2.5 Injecting Randomization

Let η be a probability distribution on $\mathbb{R}$ with mean zero and a symmetric distribution around zero. We can think of η as a noise term. Then the principal can add the noise term η independently to each x_n, $x \in N$, given any realization $x \in \mathbb{R}^N$. This generates a new probability distribution in place of π, which we denote $\pi + \eta$. The game $\Gamma_{\gamma,\eta}$ is defined exactly as before substituting $\pi + \eta$ for π. (Needless to say, more general noise — which need not be i.i.d. — can be thought of. But, to keep the analysis tractable, we refrain from doing so.) As we shall see, it can become needful for the principal to add noise $\eta \neq 0$ to ensure that $\tilde{e}$ is an NE of $\Gamma_{\gamma,\eta}$ when π is a deterministic map.

3 Disparate Students

3.1 Coarsening

We begin with the simplest example, with just two students, illustrating the benefits of coarse grading.

First suppose $N = \{\alpha, \beta\}$, i.e., there are just two students. Each student n has two effort levels, L_n (shirk) or H_n (work). The performance map is depicted in Figure 1:

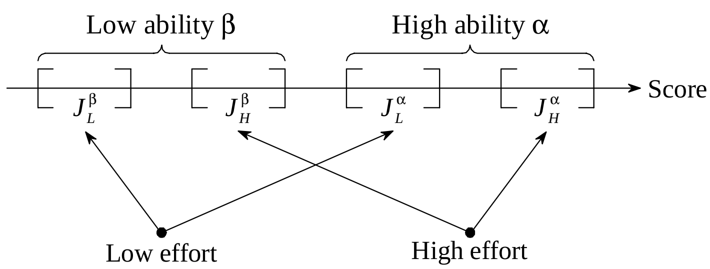

Figure 1: The Performance Map

Student n obtains marks uniformly distributed on the interval J^n_H when he works and J^n_L when he shirks. His score depends only on chance and on his own effort, and there is no interaction with the other student at this point. That happens once the professor ranks them, since each student cares about his status.

We have taken the students to have highly *disparate* abilities: $J_L^\beta < J_H^\beta < J_L^\alpha < J_H^\alpha$, i.e., α is so much more able than β, that he always comes out ahead even when he shirks and his rival works. Thus if the professor were to grade them finely, neither would work, since status could not be affected by effort. More precisely, (L_α, L_β) is the unique NE of the game[6] $\Gamma_{\gamma_{\widetilde{\mathcal{P}}}}$ where $\widetilde{\mathcal{P}} \equiv \{\{x\} : x \in \mathbb{R}\}$ denotes the finest partition — even more, it is an NE in strictly dominant strategies.

The professor can do better with a judiciously chosen coarse partition $\mathcal{P}$. Indeed consider the partition $\mathcal{P}(p) \equiv \{A, B, C\}$ shown in Figure 2. Anything below J_H^β gets grade C (including all scores in J_L^β obtained when the beta type shirks). All scores in J_H^β and J_L^α get B, as well as the bottom $(1-p)$ fraction of the scores in J_H^α. The partition is completely characterized by the single parameter $0 \le p \le 1$, specifying the fraction of J_H^α that counts for the grade A (so that we may abbreviate $\gamma_{\mathcal{P}(p)} \equiv p$, without confusion).

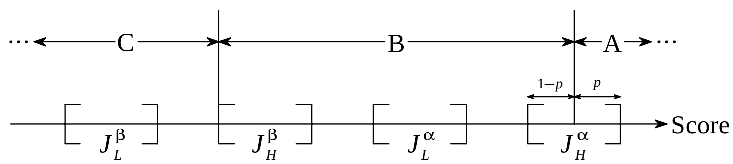

Figure 2: The partition $\mathcal{P}(p)$

The incentive $I^n(p)$ to switch from effort level L_n to H_n for any student n (assuming that his rival is working hard) is given by:

$$\begin{aligned} I^\alpha(p) &= u_p^\alpha(H_\alpha, H_\beta) - u_p^\alpha(L_\alpha, H_\beta) \\ I^\beta(p) &= u_p^\beta(H_\alpha, H_\beta) - u_p^\beta(H_\alpha, L_\beta). \end{aligned}$$

Clearly

$$u_p^\beta(H_\alpha, L_\beta) = -1$$

since, under (H_α, L_β), β always gets C and so is always behind α (who gets B with probability $1-p$ and A with probability p). Also

$$u_p^\beta(H_\alpha, H_\beta) = -p$$

since, under (H_α, H_β), α beats β with probability p. So

$$I^\beta(p) = -p - (-1) = -p + 1.$$

Similarly, one computes

$$I^\alpha(p) = p - 0 = p.$$

[6] (L_α, L_β) is in fact the strictly dominant strategy NE of $\Gamma_{\gamma_\mathcal{P}}$ for any $\mathcal{P}$ which has no cell intersecting both J_L^α and J_H^β.

Denote $d_n = H_n - L_n$, i.e., d_n is n's disutility for switching from shirk to work. Then (H_α, H_β) is a Nash equilibrium if and only if

$$I^\beta(p) = -p + 1 \geq d_\beta, \text{ i.e., } p \leq 1 - d_\beta \leq 1 \tag{1a}$$

and

$$I^\alpha(p) = p \geq d_\alpha. \tag{1b}$$

In Figure 3, we graph the two incentive functions:

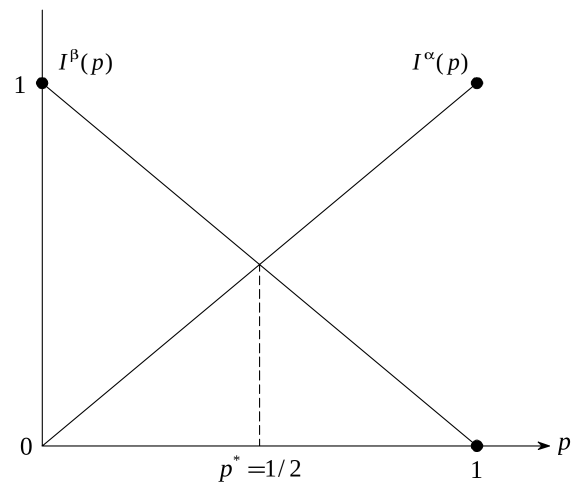

Figure 3: The Incentives $I^n(p)$

If the students have the same disutilities, i.e., $d_\alpha = d_\beta = d$, the "optimal" p is given by

$$\begin{aligned} p^* &= \arg\max_{0 \leq p \leq 1} \min\{I^\alpha(p), I^\beta(p)\} \\ &= \frac{1}{2}. \end{aligned}$$

Note that $I^n(1/2) = 1/2$ for both students n. If $d \leq 1/2$, Γ_{p^*} induces both students to work as its unique NE. Here p^* is optimal in the sense that whenever (H_α, H_β) is an NE of $\Gamma_{\gamma_\mathcal{P}}$ for some partition $\mathcal{P}$, it is also an NE of Γ_{p^*}.

Multiple Effort Levels and Less Disparateness One may imagine that each student has several effort levels and that J_L^n is the performance interval for $n \in N$ when n exerts his *second-highest* effort. Now the two students are not as heterogeneous as before: all we are postulating (in Figure 1) is that α is sufficiently more able than β so that his second-highest effort leads to uniformly better scores than β's highest effort. (The term $d_n = H_n - L_n$ must be interpreted as the extra disutility incurred when n switches from his second-highest to his highest effort.) In this setting, it is harder to sustain maximal effort as an NE (more conditions will have to be met), and our analysis gives only *necessary* conditions. Nevertheless, it shows that *if* maximal effort is an NE, it is achievable via the partitions $\mathcal{P}(p)$ that we have analyzed.

3.2 Pyramiding

Notice that the optimal grading partition, given by $p^* = 1/2$, implies:

$$\text{Expected \# of students getting } A \quad = \quad p^* = \frac{1}{2} \tag{2a}$$

$$\text{Expected \# of students getting } B \quad = \quad 1 + (1 - p^*) = \frac{3}{2}. \tag{2b}$$

In other words, optimal grading creates a pyramid with fewer A's than B's even though there are equal numbers of strong and weak students in the class.

Spence (1974) postulated that typically the weak student incurs more disutility from effort than the strong, i.e.,

$$d_\beta > d_\alpha.$$

It is evident that the Spence condition has the effect of accentuating the pyramid. Indeed, by (1), (H_α, H_β) is an NE of Γ_p iff $p \in [d_\alpha, 1 - d_\beta]$. Thus p falls as d_β rises, diminishing the expected number of A's and increasing the B's.

3.2.1 Multiple Students of Each Type

We shall show that coarsening and pyramiding persist with many students of each type. Suppose there are two students of type α, and two of type β. If a β type works, his expected exam payoff against an α worker is $p(-1)$. Since his expected exam payoff against the other β student must be zero, his total expected exam payoff if he works is then $0 + 2p(-1) = -2p$. If he shirks, he comes below all the other three for sure. His incentive to work is thus $3 - 2p$. If an α type works, his expected exam payoff against the other α student is always zero. His expected exam payoff against the two β types is $2p$. If the α type shirks, he comes equal with the β students, and behind the other α student with probability p. His incentive to work is thus $3p$. If all students n have the same disutilities d_n for effort, then the optimal partition occurs where p maximizes the minimum of $3 - 2p$ and $3p$. This occurs at $p = 3/5$. On average only 60% of α types get A.

Suppose there are N β-type students of low ability and K α-type students of high ability. All the results we obtained for $N = K = 1, 2$ remain essentially intact. The reader can check that the incentive functions become:

$$\begin{aligned} I^\beta(p) &= -pK - (-(N + K - 1)) \\ &= (-\mu^H p + 1)\delta \end{aligned}$$

where $\mu^H \equiv K/(N + K - 1)$ gives the fraction of high ability in the population, when a single low-ability student stands aside; and $\delta \equiv N + K - 1 \equiv$ utiles to a student when he beats all the others. Similarly, one can compute

$$I^\alpha(p) = \delta p.$$

It is convenient to normalize utilities, dividing incentives and effort costs by δ.

Then we may rewrite

$$\begin{aligned} d_n &= (N+K-1)^{-1} \times (\text{old } d_n) \\ I^\beta(p) &= -\mu^H p + 1 \\ I^\alpha(p) &= p \end{aligned}$$

and the analysis proceeds as before. The optimal $p = 1/(1+\mu^H)$ is obtained by solving $-\mu^H p + 1 = p$. When K and N are large and equal, μ^H is nearly 1/2, and the optimal partition p converges to 2/3. The pyramid remains. Indeed, the pyramid becomes more visible since expected # of students getting A ≈ actual # of students getting A, etc., by the law of large numbers.

We graph the incentives in Figure 4 (with $\mu^L \equiv (N-1)/(N+K-1) = 1 - \mu^H$).

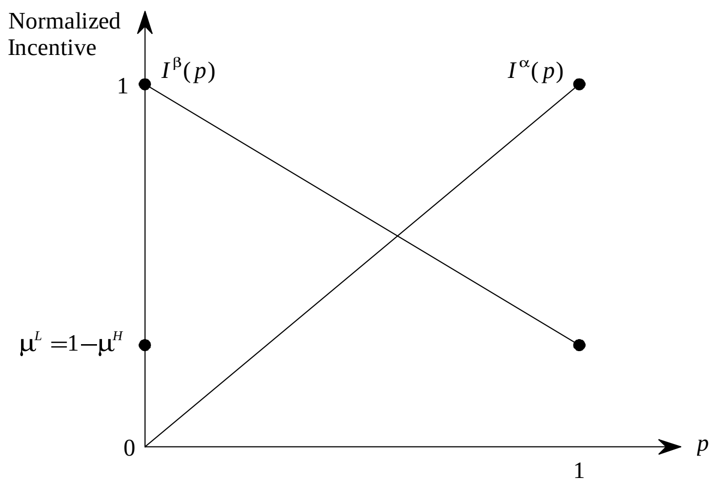

Figure 4: Normalized incentives with K α's and N β's

If the population changes to include more α-type students, this will *lower* the fraction of the α-type students who get A. (Recall that all the β-type get B.) Clearly $p = 1/(1+\mu_H)$ is decreasing in μ_H. It is interesting to observe that the proportion of A's in the whole population is always less than 1/2, since $p\mu_H = \mu_H/(1+\mu_H) \le 1/2$, with equality obtaining only when all the students are of type α (in which case $\mu_H = 1$ and $p\mu_H = 1/2$).

Observe that our solution for two disparate types is quite general. All we used is that $J_L^\beta < J_H^\beta < J_L^\alpha < J_H^\alpha$, and that the interval J_H^α has positive length, and that the exam scores of α when he works has positive density on J_H^α. In particular, all we needed was just a touch of randomness to the test scores of α when he works. Without that randomness we could *not* have created incentives for the students to work.

3.3 Many Disparate Student-Types

Our analysis carries over to any number of disparate student-types. When there are ℓ types, the optimal absolute grading partition will entail $\ell+1$ letter grades (i.e., will divide the numerical score line into $\ell + 1$ consecutive cells). Each type i will have a positive probability $0 < p_i < 1$ of obtaining grade i if he works; but will lapse into the lower grade $i - 1$ with certainty if he shirks.

We illustrate the case of three disparate types: γ (low ability), β (middle ability), α (high ability) in Figure 5.

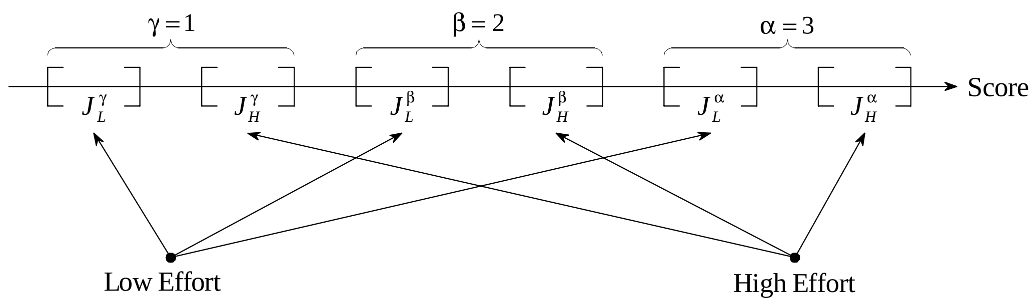

Figure 5. The Performance Map with Three Types

The partition into four grades (A, B, C, D) is depicted in Figure 6.

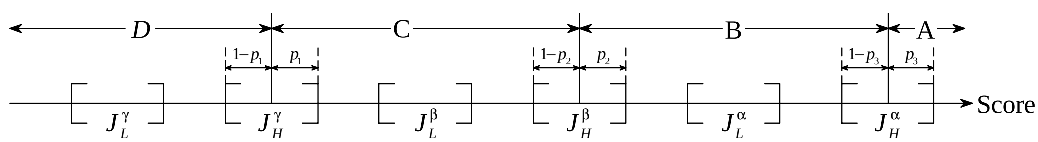

Figure 6. The Partition $\mathcal{P}(p_1, p_2, p_3)$

Let us suppose that there are N students of each type, and that N is large enough that we may take $(N-1)/N \approx 1$. (Effectively we are assuming a continuum of agents.) Then counting beating N students as one utile, the incentive to work for the ℓ types is easily computed to be

$$\begin{aligned} I^1(p) &= p_1(2 - p_2) \\ I^i(p) &= p_i(2 + (p_{i-1} - p_{i+1})) \text{ for all } 2 \leq i \leq \ell - 1 \\ I^\ell(p) &= p_\ell(1 + p_{\ell-1}). \end{aligned}$$

When working, a student of type $2 \leq i \leq \ell - 1$ might get unlucky, with probability $1 - p_i$, and find himself no better off than if he shirked. But with probability p_i he will be lucky, beating the fraction p_{i-1} of type $i - 1$ he otherwise would be equal with, and coming equal with the fraction $1 - p_{i+1}$, of type $i + 1$ he would otherwise have lost out against. In addition, he either beats (instead of equalling) or equals (instead of losing to) every student of his own type. This gives the formula $I^i(p)$ for $2 \leq i \leq \ell - 1$. Taking $p_0 = 0$ and $p_{\ell+1} = 1$ gives the formulas for $I^1(p)$ and $I^\ell(p)$.

The optimal partition p must maximize the minimum $I^i(p)$ over all $i = 1, ..., \ell$. It is immediately clear this implies $p_1 = 1$. Hence we know right off the bat that the number of students getting the lowest realized grade $i = 1$ will be greater than the population of lowest achievers (as long as $p_2 < 1$). For grades $i = 1, ..., \ell - 1$, the expected number of students obtaining grade i, when all of them work, is N multiplied by $p_i + (1 - p_{i+1})$. Precisely Np_ℓ students can expect to be awarded the top grade ℓ. Grade $i = 0$ is given only if a student of type $i = 1$ shirks.

The optimal partition can be numerically calculated for any ℓ. The pattern is roughly as follows. For reasonably large ℓ, the incentive is about 1.389 for each student. Since $p_1 = 1$, this is also the number of students (multiplied by N) receiving the lowest grade $i = 1$: $I^1(p) = p_1(2 - p_2) = (2 - p_2) = 1 + (1 - p_2)$. The top grade $i = \ell$ will be awarded to the fewest number of students, about $0.8N$. The next fewest students will be awarded the second-lowest and second-highest grades, about $0.9N$ each. Every other grade will be awarded to approximately N students.

Thus as we saw with $\ell = 2$, a small elite of A's is awarded; only 80% of alpha students ℓ get A. About 40% more $i = 1$ grades are awarded than there are very bad students. For a middling student of type i, he has a 60–70% chance of getting grade i and a 30–40% chance of getting grade $i - 1$. The number of i types graded $i - 1$ is just balanced by the number of $i + 1$ students who get graded i.

In the table below we list the optimal $(p_1, ..., p_{20})$ and the number of students (divided by N) for each grade $i = (1, ..., 20)$.

TABLE A. Pyramiding

Grade	Partition probabilities	Number of students in grade
Lowest 1	1	1.389726998
2	0.610273002	0.887761199
3	0.722511804	1.036040471
4	0.686471333	0.9887908
5	0.697680533	1.00352003
6	0.694160503	0.998897996
7	0.695262507	1.000345096
8	0.69491741	0.99989156
9	0.69502585	1.000032891
10	0.694992959	0.999986499
11	0.69500646	0.999996886
12	0.695009574	0.99997551
13	0.695034064	0.999925241
14	0.695108824	0.999760849
15	0.695347975	0.999235638
16	0.696112336	0.997568329
17	0.698544007	0.992284376
18	0.706259632	0.975830841
19	0.730428791	0.927161102
Highest 20	0.803267689	0.803267689

Observe that there are more C's than B's, and more B's than A's, but for lower grades the number of students stay equal until the very bottom is reached. The bottome grade (aside from the failing grade $i = 0$ that nobody gets) $i = 1$ is the most commonly given.

4 Homogeneous Students

Until now we have concentrated on the case where students differ substantially in their abilities. In that case, coarsening the grading allows the weaker student to compete with the stronger. We turn now to the case where all students have the same ability, and we show that coarsening still has a role to play.

4.1 An Example

Consider the situation in which N identical students take an exam. Suppose that if a student works hard, his score will be uniformly distributed on $[50\%, 100\%]$, that is, his score has density

$$f(x) = 2 \text{ if } 50\% \leq x \leq 100\%, \text{ and } 0 \text{ otherwise},$$

independent of the others' scores and effort levels. If he shirks, suppose his score has density

$$g(x) = 2x \text{ for } 0 \leq x \leq 100\%, \text{ and } 0 \text{ otherwise},$$

again independent of the others.

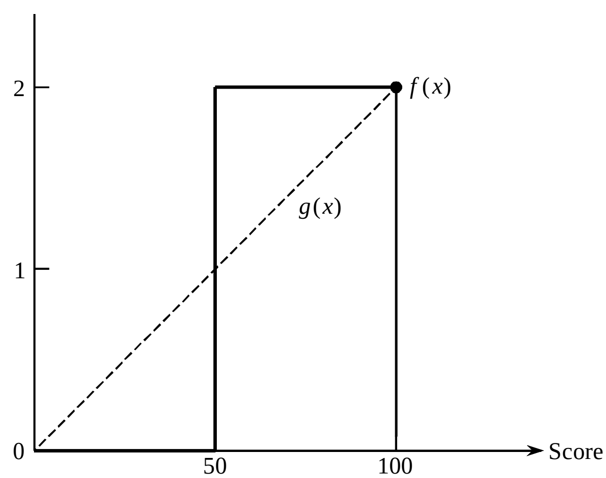

Figure 7. Score Densities

Clearly work leads to performance $f(x)$ that stochastically dominates the performance $g(x)$ from shirking on the interval $[0, 100]$. But on $[50, 100]$, indeed on any subinterval of $[50, 100]$, $g(x)$ stochastically dominates $f(x)$. (Stochastic dominance will be defined in section 4.2)

Assume all other students are working hard, and a single student is debating whether to work or shirk.

Suppose grading is perfectly fine. A student's expected exam payoff is then the sum of the probabilities he comes ahead of each other student, less the sum of the probabilities he comes behind each other student.

If he works, his expected exam payoff will be 0. This follows with no need of calculation from symmetry and the fact that *total* exam payoff is inevitably 0.

If he shirks, his expected exam payoff is equal to the probability he comes ahead of a worker, less the probability he comes behind, all multiplied by $N-1$.

The probability the shirker comes behind a worker is

$$\int_0^{1/2} 2x dx + \int_{1/2}^1 2x2(1-x)dx = \frac{1}{4} + \left[2x^2 - \frac{4}{3}x^3\right]_{1/2}^1 = \frac{7}{12}.$$

The probability the shirker comes ahead of a worker is therefore $1 - 7/12 = 5/12$, and we conclude that shirking gives an expected exam payoff

$$(N-1)\left(\frac{5}{12} - \frac{7}{12}\right) = -\frac{1}{6}(N-1).$$

This shows that the incentive to study hard is $\frac{1}{6}(N-1)$, which must be compared to the disutility of effort.

Suppose instead that just two grades are issued, namely A for scoring between 50% and 100%, and B for scoring between 0 and 50%. If a student works, along with all his $N-1$ rivals, then all will receive a score above 50% and therefore all will receive A. Each student will get a payoff of 0. If a single student fails to study, then his expected payoff is $(N-1)$ multiplied by

$$-1\int_0^{1/2} 2x dx + 0\int_{1/2}^1 2x dx = -\frac{1}{4}$$

giving an incentive to study of

$$\frac{1}{4}(N-1).$$

Since this is greater than $\frac{1}{6}(N-1)$, we see that giving only two grades creates significantly higher incentives to work than perfectly fine grading.

In fact, we will show that our partition of scores

$$\mathcal{P} = \{[0, 50\%), [50\%, 100\%]\}$$

into just two grades yields the optimal absolute grading partition.

4.2 The General Theory with iid Students

We turn now to a more general situation. Observe first that as long as the students are identical, their expected exam scores must be zero if they all work, as we noted in the example. The incentive to work therefore comes entirely from the expected payoff to a shirker who competes against hard workers. It follows that there is *no* simplification gained by assuming that each student's performance is independent of the others' effort levels. If in the example we continued to let (f, g) be the score densities of the (workers, shirker), and introduced $h \neq f$ as the score density when all work, our analysis would remain absolutely unchanged; h would be irrelevant.

On the other hand, the following independence assumption does play an important simplifying role.

Assumption: *Conditional on any choice of effort levels* $(e_1, ..., e_n)$, *students' exam scores are independent.*

We shall maintain this assumption for the rest of the paper.

Lemma 1: *Suppose two students* H *and* L *take an exam, yielding independent scores* x_H *and* x_L. *If the grading partition is* $\{(-\infty, \theta), [\theta, \infty)\}$, *then the expected exam payoff to* L *is*

$$P(x_L \in [\theta, \infty)) - P(x_H \in [\theta, \infty)).$$

Similarly, if the grading partition includes cells $[a, \theta)$, $[\theta, b)$, *for* $a < \theta < b$, *then conditional on both* x_H *and* x_L *being in* $[a, b)$, *the expected exam payoff to* L *is*

$$\frac{P(x_L \in [\theta, b))}{P(x_L \in [a, b))} - \frac{P(x_H \in [\theta, b))}{P(x_H \in [a, b))}.$$

Proof: In the first case, the expected exam payoff to L is

$$P(x_L \in [\theta, \infty) \wedge x_H \in (-\infty, \theta)) - P(x_H \in [\theta, \infty) \wedge x_L \in (-\infty, \theta)).$$

With independence, this becomes

$$\begin{aligned} & P(x_L \in [\theta, \infty))P(x_H \in (-\infty, \theta)) - P(x_H \in [\theta, \infty))P(x_L \in (-\infty, \theta)) \\ = \; & P(x_L \in [\theta, \infty))(1 - P(x_H \in [\theta, \infty)) - P(x_H \in [\theta, \infty))(1 - P(x_L \in [\theta, \infty))) \\ = \; & P(x_L \in [\theta, \infty)) - P(x_H \in [\theta, \infty)). \end{aligned}$$

The second case is analogous. ■

Corollary: *If* $\mathcal{P}$ *is a partition of scores including the cell* $[a, b)$ *and if* $\mathcal{P}^*$ *modifies* $\mathcal{P}$ *by cutting* $[a, b)$ *at* θ, *into* $[a, \theta)$ *and* $[\theta, b)$, *leaving all the other cells intact, then the move from* $\mathcal{P}$ *to* $\mathcal{P}^*$ *increases the expected exam payoff to* L *by*

$$P(x_L \in [a, b))P(x_H \in [a, b)) \left[\frac{P(x_L \in [\theta, b)}{P(x_L \in [a, b)} - \frac{P(x_H \in [\theta, b))}{P(x_H \in [a, b))} \right].$$

Proof: This follows from Lemma 1 after observing that if either $x_L \notin [a, b)$ or $x_H \notin [a, b)$, the payoff is the same under $\mathcal{P}$ or $\mathcal{P}^*$. ■

This should remind the reader of stochastic dominance.

Definition: We say that the *random variable* x *(stochastically) dominates the independent random variable* y on the interval $[a, b)$ if, whenever $P(x \in [a, b)) > 0$ and $P(y \in [a, b)) > 0$, we have

$$P(x \in [\theta, b) | x \in [a, b)) - P(y \in [\theta, b) | y \in [a, b)) \geq 0,$$

i.e.,

$$\frac{P(x \in [\theta, b))}{P(x \in [a, b))} \geq \frac{P(y \in [\theta, b))}{P(y \in [a, b))}$$

for all $\theta \in (a, b)$. In this case we write

$$x \succsim y \text{ on } [a, b).$$

Stochastic dominance has an extremely important role to play in (monotonic) grading schemes.

Lemma 2: *Suppose $x \succsim y$. Let the exam scores x and y be independent of the exam scores of every student $n = 1, ..., N-1$. Let γ be any monotonic grading scheme for N students. Then the expected exam payoff to N is at least as high under an exam score of x as it is under y.*

Proof: By independence, the payoffs from exam scores x and y depend only on their distributions. According to Theorem 1.A.1 of Shaked–Shanthikumar, there exist $\hat{x}$ and $\hat{y}$ with the same distributions as x and y respectively, such that $\hat{x} \geq \hat{y}$ with probability one. But then for *any* realization of the other $N-1$ scores, $\hat{x}$ will clearly get a (weakly) higher payoff than $\hat{y}$. ■

It will be useful to also consider a strengthened form of domination.

Definition: We say that x *uniformly dominates* y on the interval $[A, B)$ if x dominates y on every subinterval $[a, b) \subset [A, B)$. In this case we write $x \succsim_U y$ on $[A, B)$.

Both the notions of domination clearly extend to an arbitrary totally ordered set. This will be of relevance in setting up Theorems 2 and 3, where a grading partition $\mathcal{P}$ is viewed as a totally ordered set of cells.

Uniform domination can be characterized in terms of likelihood ratios in a manner that makes it much more handy to work with. But first we must restrict the random variables slightly.

Definition: We say that a random variable x is *regular* on $[a, b)$ if (a) x has a differentiable density function on $[a, b)$, or (b) x takes on a finite number of values in $[a, b)$. In either case we can speak of the density $f(t)$, for $t \in [a, b)$ or $t \in \{\alpha_1, \alpha_2, ..., \alpha_k\} \subset [a, b)$.

Lemma 3: *Let x and y be independent and regular on $[A, B)$, with density functions f and g, respectively. Then x uniformly dominates y on $[A, B)$ if and only if $f(t)/g(t)$ is increasing on $[A, B]$.*

Proof: This follows from Theorem 1.C.2 in Shaked–Shanthikumar. ■

In the leading example of this section x_L uniformly dominates x_H on $[50, 100]$.

When f and g are differentiable, $f(t)/g(t)$ is increasing if and only if $f'(t)/f(t) \geq g'(t)/g(t)$. Let $N(\mu, \sigma)$ denote the normal distribution with mean μ and standard deviation σ. If $x \sim N(\mu, \sigma)$ with density $f(t)$ and $y \sim N(\tilde{\mu}, \tilde{\sigma})$ with density $g(t)$ then

$$\frac{f'(t)}{f(t)} = \frac{-(t-\mu)}{\sigma^2}; \frac{-(t-\tilde{\mu})}{\tilde{\sigma}^2} = \frac{g'(t)}{g(t)} \quad \forall t \in (-\infty, \infty).$$

If $\mu > \tilde{\mu}$ and $\sigma = \tilde{\sigma}$, then x uniformly dominates y on all of $(-\infty, \infty)$. More generally, x will uniformly dominate y on the interval including all t such that

$$\frac{t}{\sigma^2} - \frac{t}{\tilde{\sigma}^2} < \frac{\mu}{\sigma^2} - \frac{\tilde{\mu}}{\tilde{\sigma}^2}$$

and y will uniformly dominate x on the complementary interval. Thus if $\sigma^2 < \tilde{\sigma}^2$, then x uniformly dominates y on the lower tail, and y uniformly dominates x on the upper tail.

Another instance of uniform domination occurs when an exam has k independent questions, and a student has a probability p of getting any answer correct. If another student independently has probability q of getting each question right, then the likelihood ratio condition reduces to[7]

$$\frac{p}{1-p} \geq \frac{q}{1-q}$$

As we mentioned earlier, uniform dominance can be applied to partitions. For example, let $\mathcal{P}$ be perfectly fine on $(-\infty, a)$, and on $[b, \infty)$, and let $[a, b)$ be a (masked) cell in $\mathcal{P}$. Let f and g be continuous, positive functions on $(-\infty, \infty)$. What does it mean for f/g to be increasing on $\mathcal{P}$? Define $\hat{f}(x) \equiv f(x)$ and $\hat{g}(x) \equiv g(x)$ for all $x \in (-\infty, a) \cup [b, \infty)$. Define

$$\begin{aligned} \hat{f}(x) &= \frac{1}{b-a} \int_a^b f(t)dt \text{ for } x \in [a, b) \\ \hat{g}(x) &= \frac{1}{b-a} \int_a^b g(t)dt \text{ for } x \in [a, b). \end{aligned}$$

Then $\hat{f}/\hat{g}$ must be increasing on $(-\infty, \infty)$.

We are now ready to state some theorems about the optimal absolute grading partition when students have conditionally independent scores.

The phrase "*N iid students who can work or shirk*" means that each student has two effort levels, and that assuming any one student shirks while the others work, he

[7] The notion of domination does not rely on independence. For example, suppose that with probability π the two students have chance $p_1 > q_1$ of getting each question, while with probability $1 - \pi$ they have chance $p_2 > q_2$ of getting each question; still the score of the first student would uniformly dominate that of the second. This suggests that much of our analysis can be extended to nonindependent scores, but we have not undertaken this extension here.

has score x_L while each other student k has an independent score $x_k \sim x_H$. (Here $\sim$ denotes identical distribution.) In this scenario, any grading partition creates the same incentive to work for all students, namely the negative of the expected exam score of the shirker. Thus it becomes easier to examine the general structure of optimal partitions than in the disparate case (where an increase in the incentive of one student might come at the cost of a decrease for another).

We begin by stating conditions under which a cell $[a, b)$ should not be cut by any optimal partition.

Theorem 1 (Coarseness in the Optimal Grading): *Let there be N iid students who can work or shirk. Suppose that on some interval $[a, b)$, x_L uniformly dominates x_H. Then for any partition $\mathcal{P}$ that cuts $[a, b)$, there is another partition $\mathcal{P}^*$ that gives at least as much incentive to work without cutting $[a, b)$. If x_L strictly uniformly dominates x_H on $[a, b)$, then every optimal grading partition is perfectly coarse on $[a, b)$.*

Proof: Consider the following picture:

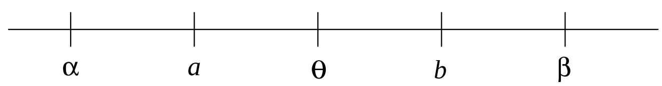

Figure 8. Cutting (a, b) at θ

Let $\mathcal{P}$ be a partition that cuts $[a, b)$ just once, that is, suppose that $\theta \in (a, b)$, and that $[\alpha, \theta) \in \mathcal{P}$ and $[\theta, \beta) \in \mathcal{P}$, where $\alpha \leq a < \theta < b \leq \beta$. (The cases where the closed end comes on the right instead of the left are handled the same way.) For the rest of the proof all probabilities will be taken conditional on x_L and x_H being in $[\alpha, \beta]$. For ease of notation, we suppress this conditionality. Thus when we write $P(x_L \in [a, b))$, we really mean $P(x_L \in [a, b))/P(x_L \in [\alpha, \beta))$, etc.

From the Corollary to Lemma 1, we know that the expected payoff to L in $\mathcal{P}$ (conditional on both x_L and x_H in $[\alpha, \beta)$) is

$$P(x_L \in [\theta, \beta)) - P(x_H \in [\theta, \beta)).$$

Suppose first that $P(x_L \in [\theta, b)) \geq P(x_H \in [\theta, b))$. Then

$$\begin{aligned} & P(x_L \in [b, \beta)) - P(x_H \in [b, \beta)) \\ \leq \quad & P(x_L \in [\theta, \beta)) - P(x_H \in [\theta, \beta)). \end{aligned}$$

It follows from the Corollary to Lemma 1 that the partition obtained from $\mathcal{P}$ by moving the cut from θ to b (i.e., by replacing $[\alpha, \theta)$ and $[\theta, \beta)$ with $[\alpha, b)$ and $[b, \beta)$) weakly lowers the expected exam payoff to L, without cutting $[a, b)$, as was to be proved.

Suppose on the other hand that $P(x_L \in [\theta, b)) < P(x_H \in [\theta, b))$. From the dominance of x_L over x_H on $[a, b)$, we conclude that $P(x_L \in [a, \theta)) < P(x_H \in [a, \theta))$. It follows immediately that

$$\begin{aligned} & P(x_L \in [a, \beta)) - P(x_H \in [a, \beta)) \\ < \quad & P(x_L \in [\theta, \beta)) - P(x_H \in [\theta, \beta)) \end{aligned}$$

showing that the partition obtained from $\mathcal{P}$ by moving the cut from θ to a (i.e., by replacing $[\alpha, \theta)$ and $[\theta, \beta)$ with $[\alpha, a)$ and $[a, \beta)$) lowers the expected payoff to L, without cutting $[a, b]$, as was to be shown.

It only remains to consider the case where $\mathcal{P}$ cuts (a, b) multiple times. If there are a finite number of cuts, that case can be reduced to the case where (a, b) is cut once, by arbitrarily choosing any cut $\theta \in (a, b)$, and then choosing the highest cut $c < \theta$ and the lowest cut $d > \theta$, and replacing (a, b) with $(a, b) \cap (c, d)$. By the *uniform* dominance of x_L over x_H on $[a, b)$, x_L dominates x_H on $[a, b) \cap [c, d)$. The same proof can then be repeated. This reduces the number of cuts by 1. The reduction can then be iterated.

Finally, suppose there is a subinterval $[c, d) \subset [a, b)$ on which $\mathcal{P}$ is perfectly fine. Change $\mathcal{P}$ to $\mathcal{P}^*$ by completely masking $[c, d)$. Since x_L dominates x_H on $[c, d)$, this masking can only (weakly) lower the expected exam payoff to L. In this way we reduce the problem to finitely many cuts. This proves the first claim of the theorem. The second claim is proved the same way. ■

Theorem 1 shows that if work leads to a normal distribution $N(\mu, \sigma)$ of scores, and shirk leads to $N(\tilde{\mu}, \tilde{\sigma})$, where $\sigma \neq \tilde{\sigma}$, then one tail of scores will be completely masked in an optimal partition.

Theorem 1 leads to a sufficient condition for the optimality of a partition $\mathcal{P}$.

Theorem 2 (Uniform Domination Implies Optimality): *Let there be N iid students who can work or shirk. Let $\mathcal{P}$ be a partition such that for every cell $[a, b) \in \mathcal{P}$, $x_L \succsim_U x_H$ on $[a, b)$, and such that $x_H \succsim_U x_L$ on the totally ordered set $\mathcal{P}$. Then $\mathcal{P}$ is optimal.*

Proof: Suppose $\mathcal{P}'$ does better than $\mathcal{P}$. From Theorem 1, we know that there is $\mathcal{P}''$ that does at least as well as $\mathcal{P}'$, and which does not cut any cell in $\mathcal{P}$. But every cell in $\mathcal{P}''$ is refined by cells in $\mathcal{P}$. Since $x_H \succsim_U x_L$ on $\mathcal{P}$, it follows that the payoff to x_L is weakly lower in $\mathcal{P}$ than in $\mathcal{P}''$, a contradiction. ■

Suppose that all students are *ex ante* identical, with independent, and normally distributed exam scores. If hard work raises a student's expected exam score, without changing its variance, then Theorem 2 implies that an optimal grading scheme is to post the exact scores.

Similarly, if the K exam questions are identical, independent trials, and if hard work allows a student to raise his probability of getting each answer right, then again an optimal grading scheme is to reveal the exact scores.

But consider the leading example of this section. There we found that giving just two grades, A and B, improved incentives beyond what could be achieved by fully revealing the scores. Theorem 2 guarantees that this is indeed an optimal partition. Inside the cell $[0, 50)$, x_H has probability zero, so x_L trivially uniformly dominates it. Inside the other cell $[50, 100)$, $f(t)/g(t) = 2/2t = 1/t$ is strictly falling, so by Lemma 3, x_L uniformly dominates x_H. Across cells we can check that x_H uniformly dominates x_L. On $[0, 50)$, we can define the effective density of a worker as $\hat{f}(t) = 0$, and that of a shirker as $\hat{g}(t) = .5$. On $[50, 100)$ the effective densities become $\hat{f}(t) = 2$ and $\hat{g}(t) = 1.5$. Clearly $\hat{f}(t)/\hat{g}(t)$ is increasing. Now we can apply Theorem 2.

In our next theorem we describe *necessary* conditions for a partition to be optimal, when agents are homogeneous. The "outside" condition, that $x_H \succsim_U x_L$ on $\mathcal{P}$, is the same as the sufficient "outside" condition appearing in Theorem 2. But the "inside" condition $x_L \succsim x_H$ on each cell $[a, b)$ in $\mathcal{P}$ is weaker than the sufficient "inside" condition $x_L \succsim_U x_H$ appearing in Theorem 2.

Theorem 3 (Optimality Implies Domination): *Let there be N iid students who can work or shirk. Let $\mathcal{P}$ be an optimal absolute grading partition. Then for any cell $[a, b) \in \mathcal{P}$, $x_L \succsim x_H$ on $[a, b)$. Furthermore, if exam performances are regular, then $x_H \succsim_U x_L$ on the totally ordered set $\mathcal{P}$.*

Proof: Consider any cell $[a, b)$ in $\mathcal{P}$ such that $P(x_L \in [a, b))P(x_H \in [a, b)) > 0$. Suppose there is some $\theta \in [a, b)$ with

$$\frac{P(x_L \in [\theta, b))}{P(x_L \in [a, b))} - \frac{P(x_H \in [\theta, b))}{P(x_H \in [a, b))} < 0.$$

Change $\mathcal{P}$ to $\mathcal{P}^*$ by replacing $[a, b)$ with $[a, \theta)$ and $[\theta, b)$. By the Corollary to Lemma 1, this must lower the expected exam payoff to the shirker against each worker. But this means that $\mathcal{P}^*$ is a better partition than $\mathcal{P}$, a contradiction proving that $x_L \succsim x_H$ on $[a, b)$.

Suppose now that the performances are regular. Consider two consecutive cells $[a, b)$ and $[b, c)$ in $\mathcal{P}$ whose union $(a, b] \cup (b, c]$ has positive probability of being reached by both x_L and x_H. Then it is clear from the Corollary to Lemma 1 that

$$\frac{P(x_H \in [b, c))}{P(x_H \in [a, c))} \geq \frac{P(x_L \in [b, c))}{P(x_L \in [a, c))},$$

otherwise the partition $\mathcal{P}^*$ obtained from $\mathcal{P}$ by replacing the two cells $[a, b)$ and $[b, c)$ with the single cell $[a, c)$ would lower the expected exam score to L, contradicting the optimality of $\mathcal{P}$.

First consider the case where the exam score space is discrete. Without loss of generality, we can drop any scores that have probability zero of being reached by both x_L and x_H. Also we can then lump together into a single cell *consecutive* exam scores which x_L reaches with zero probability. The same can thereafter be done for consecutive scores that x_H reaches with zero probability. Now the argument of the previous paragraph shows that the likelihood ratio property must hold for the cells of $\mathcal{P}$ proving (by Lemma 3) that $x_H \succsim_U x_L$ on $\mathcal{P}$ in the discrete case.

As for the continuous case, the partition $\mathcal{P}$ must consist of discrete intervals, each of which is completely fine or completely masked. Thinking of the masked intervals as single points, the proof proceeds exactly as above for consecutive masked intervals.

It remains to consider an interval $[a, b)$ on which $\mathcal{P}$ is completely fine. Let f_H be the (differentiable) density for x_H, and let f_L be the (differentiable) density for x_L. We claim f_H/f_L must have nonnegative derivative, and thus be weakly increasing, on all of $[a, b)$. If the derivative is negative somewhere in $[a, b)$ then, on some small subinterval $(\alpha, \beta) \subset [a, b)$, f_H/f_L must be strictly decreasing. But that immediately implies that conditional on $x_H \in [\alpha, \beta)$ and $x_L \in [\alpha, \beta)$, x_L is more likely to come ahead of x_H than behind. Modifying $\mathcal{P}$ to $\mathcal{P}^*$ by completely masking $[\alpha, \beta)$ gives (by the Corollary to Lemma 1) a lower expected exam payoff to L, contradicting the optimality of $\mathcal{P}$. Thus the likelihood ratio properly holds on completely fine intervals.

Finally, for each masked interval $(c, d]$ in $\mathcal{P}$, consider the pseudo density $\hat{f}_H = P(x_H \in [c, d))/(d - c)$ and $\hat{f}_L = P(x_L \in [c, d))/(d - c)$. We shall now show that at any cut θ between a completely fine interval $[a, \theta)$ and a completely masked interval $[\theta, b)$ in $\mathcal{P}$, $f_H(\theta)/f_L(\theta) \leq \hat{f}_H(\theta)/\hat{f}_L(\theta)$. If not, we shall show that the partition $\mathcal{P}^*$ obtained from $\mathcal{P}$ by slightly expanding the masked interval $[\theta, b)$ to $[\theta - \varepsilon, b)$ lowers the expected exam payoff to L.

Indeed, the change in expected payoff to L is

$$\begin{aligned}
&P(x_H \in [\theta, b))P(\theta - \varepsilon \leq x_L < \theta) - P(x_L \in [\theta, b))P(\theta - \varepsilon \leq x_H < \theta) \\
&\quad + P(\theta - \varepsilon \leq x_H < \theta)P(\theta - \varepsilon \leq x_L < \theta)[P(x_H > x_L | \theta - \varepsilon \leq x_L, x_H < \theta) \\
&\quad - P(x_L > x_H | \theta - \varepsilon \leq x_L, x_H < \theta)].
\end{aligned}$$

Observe that the third term goes to zero as ε^2 when $\varepsilon \to 0$, whereas the first two terms are of the order of ε. As $\varepsilon \to 0$, $P(\theta - \varepsilon \leq x_H < \theta)$ converges to $\varepsilon f_H(\theta)$, and $P(\theta - \varepsilon \leq x_L < \theta)$ converges to $\varepsilon f_L(\theta)$. Thus if $f_H(\theta)/f_L(\theta) > \hat{f}_H(\theta)/\hat{f}_L(\theta)$, then the first two terms add to less than zero. This shows that the extra masking obtained by lowering the cut θ to $\theta - \varepsilon$ reduces the expected exam payoff to L, contradicting the optimality of $\mathcal{P}$.

Thus we have proved that the pseudo densities $\tilde{f}_H$ and $\tilde{f}_L$ (defined by f on perfectly fine intervals in $\mathcal{P}$ and by $\hat{f}$ for masked intervals) satisfy the likelihood ratio dominance of $\tilde{f}_H$ over $\tilde{f}_L$. Applying Lemma 3 proves the theorem. ■

5 Heterogeneous Students

Consider now the situation in which students are neither identical nor disparate. To be concrete, suppose that an exam consists of K questions. Each student $n = 1, ..., N$ who works has a probability p_n of getting any question right, where answers are independent across students and questions. Suppose that if n shirks this probability drops to p_{n-1} where $0 = p_0 < p_1 < p_2 < \cdots < p_N = 1$, and $p_n - p_{n-1} = 1/N$ for all n.

We have seen that working gives each student an exam performance that uniformly dominates his performance from shirking. Were all the students identical (say $p_n(\text{work}) = 1/2$ for all n, and $p_n(\text{shirk}) < 1/2$ for all n) then the optimal grading partition would be perfectly fine, by Theorem 2. But on account of the heterogeneity, each student must compete with the performance of other students who are not like him.

A standard variant of the central limit theorem shows that as N gets large, the *class* performance converges to the distribution given by $p = 1/2$. For p_n near 0 (and p_n near 1), a student n will almost surely finish near the bottom (near the top) whether or not he works. Thus with perfectly fine grading the best and worst students have little incentive to work. The interesting thing is that coarse grading will increase their incentive to work. Since students in the middle with p_n near $1/2$ can surpass a large number of others by switching from shirk to work, they already have huge incentives to work. Even if coarse grading diminishes the middling students' incentives, it is still the most effective device to incentivize *all* students to work, i.e., to maximize the minimum incentive.

We illustrate this by considering the case where $K = 26$ and $N = 10$. We simulate the situation with $2^{26} = 67,108,864$ students. In Table B we report the expected exam payoffs for ability levels $p = .1, .2, ..., .9$, who work, when scores are partitioned into grades $\{[0, g), [g, 2g), ..., [rg, 27)\}$, where $rg \leq 26 < (r+1)g$, for $g = 1, ..., 9$. Higher g corresponds to coarser partitions. For $g = 1$ we get the perfectly fine partition where all scores are reported. For $g = 9$, the grading partition corresponds to just three grades $\{[0, 9), [9, 18), [18, 27)\}$. Note that low-ability types get very low expected exam payoffs, losing to nearly everyone in the class.

In Table C we give the incentives to work, obtained from Table B by subtracting expected exam payoffs for p_{n-1} from expected exam payoffs for p_n. Notice how coarsening the partition helps all the very bad and very good students.

Finally, in Table D we compare the incentives from the partition $g = 9$ that gives just three grades to the fine partition $g = 1$ that reports scores exactly. Everybody except the middling students $p = .5$ and $p = .6$ have higher incentives to work from coarse grading. The maximin is also higher for $g = 9$ than for $g = 1$.

TABLE B. Expected Exam Payoffs

	Size of partition cells								
	1	2	3	4	5	6	7	8	9
0.1	-67060341.99	-67027557.41	-66954572.62	-66798284.52	-66452484.31	-66279793.62	-66334980.98	-65969700.87	-64533683.04
0.2	-65768474.36	-65404074.77	-64749982.27	-63751432.39	-62476223.82	-58974576.83	-56995198.89	-58766361.91	-60749201.08
0.3	-57811907.91	-56765291.46	-55061210.25	-52705913.55	-50553306.28	-47476806.64	-38901651.72	-35793564.73	-40511823.36
0.4	-35591919.04	-34470982.35	-32739333.56	-30669192.43	-27533025.36	-27565352.02	-22987902.24	-15390212.32	-14405734.5
0.5	0	0	0	0	0	0	0	0	0
0.6	35591919.04	34470982.29	32739333.56	30546272.77	28812711.55	23743732.17	25973646.66	24445855.03	14405734.5
0.7	57811907.91	56765291.32	55061210.25	52885114.95	49945236.31	46357223.89	41579600.03	47925472.56	40511823.36
0.8	65768474.36	65404074.52	64749982.27	63667698.85	61935421.24	61917720.25	55004931.85	56537217.42	60749201.08
0.9	67060341.99	67027557.12	66954572.62	66823769.13	66651744.43	65837255.16	65894986.17	61737403.64	64533683.04

TABLE C. Incentives

High	Size of partition cells								
ability	1	2	3	4	5	6	7	8	9
0.2	1291868	1623483	2204590	3046852	3976260	7305217	9339782	7203339	3784482
0.3	7956566	8638783	9688772	11045519	11922918	11497770	18093547	22972797	20237378
0.4	22219989	22294309	22321877	22036721	23020281	19911455	15913749	20403352	26106089
0.5	35591919	34470982	32739334	30669192	27533025	27565352	22987902	15390212	14405735
0.6	35591919	34470982	32739334	30546273	28812712	23743732	25973647	24445855	14405735
0.7	22219989	22294309	22321877	22338842	21132525	22613492	15605953	23479618	26106089
0.8	7956566	8638783	9688772	10782584	11990185	15560496	13425332	8611745	20237378
0.9	1291868	1623483	2204590	3156070	4716323	3919535	10890054	5200186	3784482

TABLE D. Advantage of Coarse Grading

	Incentives from g = 9	Incentives from fine partition g = 1
p = .2	3784481.96	1291867.625
p = .3	20237377.72	7956566.455
p = .4	26106088.86	22219988.87
p = .5	14405734.5	35591919.04
p = .6	14405734.5	35591919.04
p = .7	26106088.86	22219988.87
p = .8	20237377.72	7956566.455
p = .9	3784481.96	1291867.625

6 Grading on a Curve

We have assumed that students care only about their relative grade. It would seem therefore that relative grading, i.e., grading on a curve, would provide the best incentives. But in fact the contrary is true. We shall prove that when all the students are *ex ante* identical, it is always better to grade according to an absolute scale.

Before turning to the case of homogeneous students, recall the model of disparate students described in Section 3. With one weak student and one strong student, grading on a curve provides no incentive whatsoever. (Either both get A, or else the strong always gets A and the weak always gets B.) When there are two α students and two β students, it is clear that the optimal way to grade on a curves is to declare the top student an A, the next two a B and the last a C. This generates incentives $3/2 = (1/2)3 + (1/2)0 - 0$ for each α-student and $3/2 = (1/2)0 + (1/2)(-3) - (-3)$ for each β-student.

Thus absolute grading strictly dominates grading on a curve, since we showed earlier in Section 3.1.1 that absolute grading with $p^* = 3/5$ gives incentives of $9/5$ to every student. It is also worth noting that, even if one were restricted to grading on a curve, coarse grading (as above) would be better than fine, since fine generates only the incentive of 1 for both student-types.

Consider our leading example of homogeneous students from Section 4, in which a hard working student will get a score uniformly distributed between 50 and 100, and the density of a shirker's score is $2x$. The optimal absolute grading system is to give A for scores in $[50, 100)$ and B for scores in $(0, 50)$. With three students we computed that the incentive to work is $(3 - 1)(1/4) = 1/2$.

Now consider grading on a curve. Giving everybody an A provides no incentive at all. Giving three grades is just like the perfectly fine partition with absolute grading and is therefore not as good as the optimal partition $([0, 50), [50, 100))$. (Indeed we computed that fine grading gave incentive 1/3).

It can be checked that with a curve that gives two B's and one A, the incentive to work is $99/324 < 1/2$, while with a curve that gives one B and two A's it is $63/324 < 1/2$.

None of the possible grading curves yields incentives as high as absolute grading. We now establish the same conclusion in the general case of N independent, *ex ante* identical students. But first we establish three more lemmas.

Lemma 4: *Denote scores in $[\theta, \infty)$ as A. Suppose $N-1$ students work hard, and each has probability p_G of getting an A, while one student shirks and has probability $p_B \leq p_G$ of getting an A. Suppose all scores are independent. If exactly K students wind up with A, the conditional probability that the shirker got A is at most K/N, while the probability any hard worker got A is at least K/N.*

Proof: The conditional probability the shirker got A is

$$\frac{p_B \begin{pmatrix} N-1 \\ K-1 \end{pmatrix} p_G^{K-1}(1-p_G)^{N-K}}{p_B \begin{pmatrix} N-1 \\ K-1 \end{pmatrix} p_G^{K-1}(1-p_G)^{N-K} + (1-p_B) \begin{pmatrix} N-1 \\ K \end{pmatrix} p_G^{K}(1-p_G)^{N-K-1}}$$

which is strictly monotonically increasing in p_B (as can easily be seen by dividing numerator and denominator by $(1-p_B)$). But when $p_B = p_G$, symmetry implies that the expression must be exactly K/N. Hence the probability the shirker got A is at most K/N. Since exactly the proportion K/N students did get A, the probability of the good students getting A must then be at least K/N. ■

Lemma 5: *Suppose $N-1$ students work hard, and have i.i.d. probabilities $p = (p_A, p_B, ..., p_Z)$ of getting absolute grades $A, B, ..., Z$, while one shirker has independent probabilities $q = (q_A, q_B, ..., q_Z)$. Suppose p uniformly dominates q. If exactly $\mu = (\mu_A, \mu_B, ..., \mu_Z)$ students wind up with grades $(A, B, ..., Z)$, then the posterior probability distribution $\hat{p} = (\hat{p}_A, \hat{p}_B, ..., \hat{p}_Z)$ of a hard working student getting grades $A, B, ..., Z$ uniformly dominates the posterior probabilities $\hat{q} = (\hat{q}_A, \hat{q}_B, ..., \hat{q}_Z)$ of the shirker.*

Proof: Simply apply the previous lemma repeatedly. A partition into n cells can always be obtained by repeated cuts of cells into z pieces. ■

Lemma 6: *Let there be N iid students who can work or shirk. Suppose P^* is an optimal absolute grading partition, giving the highest possible incentive to work. Suppose one student shirks while the others work hard. Then conditional on any distribution of absolute exam grades given by P^*, the expected exam payoff to the shirker is at most zero.*

Proof: If P^* is optimal, then by Lemma 1, the *ex ante* probability $p = (p_A, p_B, ..., p_Z)$ of a hard working student getting grades $A, B, ..., Z$ must uniformly dominates the *ex ante* probabilities $q = (q_A, q_B, ..., q_Z)$ of the shirker. By the previous lemma, the posterior probability $\hat{p}$ uniformly dominates $\hat{q}$. Since exam payoffs are monotonic in the grade, the posterior expected exam payoff is at least as high for a hard worker as for a shirker. Since total exam payoffs are always zero, the posterior expected exam payoff of the shirker plus $(N-1)$ times the posterior expected exam payoff of a hard worker must be zero. The conclusion follows. ∎

Theorem 4: *Let there be N iid students who can work or shirk. Let all their scores be regular. Then optimal absolute grading gives at least as much incentive for hard work as any grading on a curve.*

Proof: Let $\mathcal{P}^*$ be an optimal absolute grading partition of $(-\infty, \infty)$. Let $\mathbb{Q}$ be any partition of class rank $\{1, 2, ..., |N|\}$, representing an arbitrary grading on a curve.

For any possible exam scores $x = (x_n)_{n \in N}$, we can specify an absolute grade distribution by counting the number of students who get A's, B's, and so on. Call this the absolute grade distribution $\Delta(x)$ generated by x. It suffices to show that conditional on *any* grade distribution Δ, the expected exam payoff to the shirker from absolute grading is no higher than his expected payoff from grading on the curve Q. A picture helps to clarify the situation.

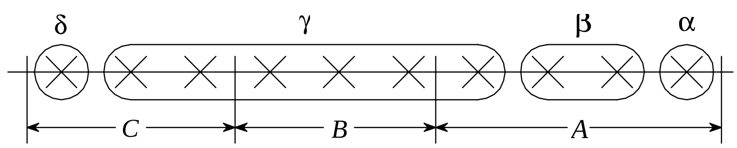

Figure 9. Absolute vs. Relative Grading

In the picture there are 4 A's, 3 B's, and 3 C's on the absolute scale. The relative scale gives grade α to the top score, β to scores 2 and 3, γ to $\{4, 5, 6, 7, 8, 9\}$, and δ to the 10th highest score. We must show that the expected score of the shirker is worse under the absolute scale P^*, than under the relative scale Q, given that according to P^* there were 4 A's, 3 B's, and 3 C's awarded.

Conditional on the absolute grade distribution Δ, let $\hat{p}$ be the posterior probability that a worker receives each grade, and let $\hat{q}$ be the posterior probability that the shirker receives each grade. From Theorem 3, Lemma 3, and Lemma 5, $\hat{p}/\hat{q}$ is increasing as we go up the cells of the optimal partition $\mathcal{P}^*$.

Define the join $P^* \vee Q$ of $\mathcal{P}^*$ and $\mathbb{Q}$ as follows: given scores $(x_n)_{n \in N}$, the exam grade for x_n is strictly higher according to $P^* \vee Q$ than the exam grade for x_m iff *either* the absolute letter grade for x_n is strictly higher than for x_m, *or* the relative grade for x_n is strictly higher than for x_m. In the diagram that is achieved by cutting the grade cell γ into three grades (one γ^+, three γ, and two γ^-) where it is intersected by P^*.

We will now argue that if we grade according to the join $\mathcal{P}^* \vee \mathbb{Q}$ then the expected exam payoff of the shirker is no more than it was in Q. If a relative grade, such as γ

in Q, is refined in $P^* \vee Q$, we show that the expected score of the shirker, conditional on being in γ to begin with, must go down. Since splitting γ does not affect scores against students outside γ, it suffices to show that the expected exam score of the shirker against the other students in γ must be less than zero.

Let $\tilde{p}$ (or $\tilde{q}$) be the probabilities of a worker (or a shirker) getting each absolute grade, conditional on the grade distribution Δ and the worker (or the shirker) being in γ. If a partition cell (like B in the example) is completely included in γ, then

$$\frac{\tilde{p}_B}{\tilde{q}_B} = \frac{\hat{p}_B}{\hat{q}_B} \cdot \frac{P[x_L \in \gamma|\Delta]}{P[x_H \in \gamma|\Delta]}.$$

Consider the bottom absolute grade that intersects γ (namely C in the example). Because the shirker stochastically dominates any worker, conditional on them both getting C, and since $\gamma \cap C$ includes the top tail of the C distribution,

$$\frac{\tilde{p}_C}{\tilde{q}_C} \leq \frac{\hat{p}_C}{\hat{q}_C} \cdot \frac{P[x_L \in \gamma|\Delta]}{P[x_H \in \gamma|\Delta]}.$$

Consider the top absolute grade that intersects γ, namely A in the example. Since $\gamma \cap A$ includes the bottom tail of the A distribution,

$$\frac{\tilde{p}_A}{\tilde{q}_A} \geq \frac{\hat{p}_A}{\hat{q}_A} \cdot \frac{P[x_L \in \gamma|\Delta]}{P[x_H \in \gamma|\Delta]}.$$

Since $\hat{p}_C/\hat{q}_C \leq \hat{p}_B/\hat{q}_B \leq \hat{p}_A/\hat{q}_A$, it follows that $\tilde{p}_C/\tilde{q}_C \leq \tilde{p}_B/\tilde{q}_B \leq \tilde{p}_A/\tilde{q}_A$. Hence $\tilde{p}$ uniformly dominates $\tilde{q}$, and so the payoff to the shirker from $\mathcal{P}^* \vee \mathbb{Q}$ must be worse than zero against a worker, conditional on Δ and the shirker being in γ.

To conclude the proof, we need only show that the expected exam payoff of the shirker in P^* is even lower (weakly) than his expected exam payoff in $P^* \vee Q$. This follows at once from the fact (see Theorem 1) that conditional on being in a cell of P^*, the score of the shirker dominates the score of a worker. But then, by Lemma 2, any monotonic grading scheme within cells of $\mathcal{P}^*$ (as happens in $P^* \vee Q$) will weakly increase the payoff of the shirker. ■

7 Midterms

The introduction of midterms before a final makes it even more necessary to have coarse grading.

The point can be easily seen via our simple example (see Section 3.1) of two disparate students. Suppose that the final has been assigned a weight of $0 < \lambda < 1$, since it covers a fraction λ of the course. Then it is natural to assume that if a student n works for the final, he incurs disutility $\lambda d_n \equiv \lambda(H_n - L_n)$ when he switches from shirk to work. The weight on the midterm is $1 - \lambda$ with associated disutility $(1 - \lambda)d_n$.

We imagine a two-period game in extensive form between the students. In period 1, both simultaneously choose effort levels (work or shirk) for the midterm. Their performance is as in Figure 1. The scores of the midterm are made public, i.e., each student finds out not only his own, but his rival's score. Then they enter period 2 and again simultaneously choose effort levels for the final. The performance is once more given by Figure 1. (It is independent of past history and depends only on the current effort level.) Suppose students care only about their (relative) grade for the course. Suppose also that the professor wants them to work for both exams. How should the two exam grades be combined into a course grade?

One common practice would be simply to add the midterm score (multiplied by $1-\lambda$) to the final exam score (multiplied by λ), and then to assign a letter grade to the aggregate score in accordance with a partition $\mathcal{P}(p)$ (see Figure 2).

We shall see that with such a grading scheme, it is often *not* feasible to incentivize both students to work on both exam under all circumstances (i.e., at all information sets of the game tree).

But if the midterm is also just given a letter grade, interpreted as some representative score out of the cell it stood for, then it becomes much easier to incentivize them.

Let us turn to a precise calculation. Denote $J_L^n = [a_L^n, b_L^n]$ and $J_H^n = [a_H^n, b_H^n]$ for $n = \alpha, \beta$ in Figure 1. Assume that $\lambda b_L^n + (1-\lambda)b_H^n < a_H^n$ and $(1-\lambda)b_L^n + \lambda b_H^n < a_H^n$ for $n = \alpha, \beta$, i.e., the intervals J_L^n, J_H^n are sufficiently separated so that, if a student shirks in either exam, his weighted average cannot be in his upper interval.

Suppose the course grade is obtained by averaging the midterm score (weighted by $1-\lambda$) and the final score (weighted by λ), and then assigning a letter grade according to the old partition $\mathcal{P}(p)$.

Suppose both students have worked in period 1 and have entered period 2 after observing the midterm numerical scores.

Suppose the good student α was lucky on the midterm, scoring the maximum b_H^α. He will get an A for the course if his score on the final x satisfies

$$(1-\lambda)b_H^\alpha + \lambda x \geq (1-p)b_H^\alpha + p a_H^\alpha.$$

If $\lambda \leq p$, this will always hold, provided that α works on the final, and α will certainly get an A for the course. But then, realizing that he can never come equal with α, β will not bother to work at all studying for the final. If $\lambda > p$, then the probability that α gets an A is p/λ, since the above inequality is equivalent to $x \geq b_H^\alpha - \frac{p}{\lambda}[b_H^\alpha - a_H^\alpha]$. Hence β's incentive to study for the final (which guarantees him a B) is $(1-\frac{p}{\lambda}) + \frac{p}{\lambda}(-1) - (-1) = 1 - \frac{p}{\lambda}$. This must exceed his disutility λd_β, so we obtain:

$$p \leq \lambda - \lambda^2 d_\beta. \tag{4a}$$

In the unlucky midterm scenario for α, when he gets a_H^α, we must have $\lambda > 1-p$, otherwise α will abandon work on the final. Then it is easy to compute that α's incentive to work on the final is $1 - 1/\lambda + p/\lambda$ which must exceed λd_α, giving us

$$p \geq \lambda^2 d_\alpha + 1 - \lambda. \tag{4b}$$

(Note that, since $0 < \lambda < 1$, (4a) implies $\lambda > p$ and (4b) implies $\lambda > 1-p$.) To sum up: for both students to work for the final exam, no matter how the midterms turned out, it is necessary (and sufficient) that

$$p \in [\lambda^2 d_\alpha + 1 - \lambda, \lambda - \lambda^2 d_\beta]. \tag{5}$$

(If we take the open interval, they will strictly want to work.)

Clearly there is no solution to (5) if $\lambda \leq 1/2$. This seems to have a counterpart in the real world, since it is quite uncommon to see a final which carries less weight than the midterm.

Indeed if $d_\alpha = d_\beta \equiv d$, then (5) has a solution iff

$$2\lambda^2 d + 1 < 2\lambda \tag{6}$$

showing the need to weight the final sufficiently more than the midterm, in order to keep students from shirking after the midterm.

Take $\lambda = 0.6$ and $d = 0.3$. Then, by (6), it is impossible to incentivize the students to work if midterm numerical scores are revealed.

Suppose, however, that information is coarsened at this interim stage and only letter grades are assigned in the midterm, in accordance with $\mathcal{P}(p)$ for $p = 1/2$, as in Figure 2. Furthermore suppose it is announced that, for the purposes of computing the final weighted average, an A in the midterm will be viewed as a score of $(a_H^\alpha + b_H^\alpha)/2$ and a B as $\frac{3}{4}a_H^\alpha + \frac{1}{4}b_H^\alpha$ (and C as a_L^β). We submit[8] that in this scenario both students will be driven to work (and leave it to the reader to check it).

8 Games of Status in Other Settings: An Application to the Optimal Wage Structure

Status-consciousness is widely prevalent in real life though it has not found a place in much of economic theory. The fact that people (unlike Robinson Crusoe) care not only about their direct consumption, but also about *relative* consumption vis-á-vis others of their social group, in particular about the higher status that superior consumption brings with it, can have profound implications for their behavior. We illustrate the phenomenon with a simple example of wage setting.

Consider a principal who wishes to hire two agents α and β to work for him. As before α is more skilled than β: indeed their performance is as in Figure 1, with the exam score now reinterpreted as output produced. The principal's problem is to announce wages as a (weakly) monotonic function of output so as to induce both agents to work hard, while minimizing his total wage bill. For simplicity we suppose that not only the principal, but also the agents, are risk-neutral. Each agent is moreover concerned about whether his wage exceeds that of the other, i.e., about his status. Thus if α and β are paid x_α and x_β respectively, then $n \in \{\alpha, \beta\}$ obtains utility

$$u^n(x_n, x_m) = x_n + \tau \ \mathrm{sgn}\{x_n - x_m\}$$

[8] Now assume, in addition to sufficient separation, that $\frac{1}{2}a_L^\beta + \frac{1}{2}a_H^\alpha < a_L^\beta$.

where $m \equiv \{\alpha, \beta\} \backslash \{n\}$ denotes the rival of n; $\text{sgn}\{x\}$ is the sign of x (i.e., $\text{sgn}\{x\} = 1$ if $x > 0$, $\text{sgn}\{x\} = 0$ if $x = 0$, $\text{sgn}\{x\} = -1$ if $x < 0$); and τ is a positive parameter that measures the importance of status.

If, as in traditional analysis, wage were to be commensurate with output (e.g., the "piece-rate" contract wherein each agent is paid a fixed percentage of what he produces) then β would always obtain lower status than α, and the agents would work solely for the direct consumption of wages. By making wages sufficiently high, it is evident that the principal can lure both to work. But he can do much better by designing a wage function that consists of a few grades, pooling J_H^β and the lower half of J_H^α into one wage-grade (much like the grade B in the example of Section 3.1). For this has the effect of creating a game of status between the agents, so that each now strives not just for wages but also for the status conferred on him when his wage overtakes that of his rival. Grade-wages bring status into play and unleash competition which stands to the benefit of the principal.

To be precise, consider a standard wage function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ which is *strictly* monotonic. Denote $\int_{J_k^n} \psi(t)dt \equiv w_k^n$ for $n \in \{\alpha, \beta\}$ and $k \in \{H, L\}$. It is easy to check that, if β shirks, his (expected wage) payoff is $w_L^\beta - \tau$; if he works, it is $w_H^\beta - \tau$; and so β has incentive $w_H^\beta - w_L^\beta$ to switch from shirk to work. Since this must exceed d_β, we must have $w_H^\beta - w_L^\beta \geq d_\beta$. But clearly the principal should announce wages as low as he can over the interval J_L^β (zero ideally, except that we have constrained him to make ψ strictly monotonic). So, taking $w_L^\beta = \varepsilon \approx 0$, we may rewrite the incentive condition for β as

$$w_H^\beta \geq d_\beta + \varepsilon. \tag{7}$$

Similarly, the incentive condition for α is $w_H^\alpha - w_L^\alpha \geq d_\alpha$ which, by (7) and the fact that $w_L^\alpha > w_H^\beta$ on account of the strict monotonicity of ψ, yields

$$w_H^\alpha > d_\alpha + d_\beta + \varepsilon. \tag{8}$$

From (7) and (8) we get a lower bound on the principal's wage bill in the traditional case

$$w_H^\alpha + w_H^\beta > d_\alpha + 2d_\beta + 2\varepsilon. \tag{9}$$

Consider instead a wage function that partitions the output space into $\mathcal{P}(p)$ as in Figure 2, assigning constant wages v_0, v_1, v_2 in lieu of the grades A, B, C with (of course) $0 = v_0 < v_1 < v_2$. In this scenario, β obtains[9] (expected wage) payoff $-\tau$ if he shirks and $v_1 - p\tau$ if he works, yielding the incentive condition

$$v_1 \geq d_\beta - (1-p)\tau. \tag{10}$$

Similarly α obtains payoffs v_1 and $(1-p)v_1 + p(v_2 + \tau)$ from shirking and working, yielding the incentive condition

$$pv_2 \geq d_\alpha + pv_1 - p\tau. \tag{11}$$

[9] As usual, when an agent contemplates switching between shirk and work, his rival is assumed to be working throughout.

The wage bill incurred by the principal is $v_1 + (1-p)v_1 + pv_2 = (2-p)v_1 + pv_2$. The principal must minimize this subject to $0 \le p \le 1$ and $0 < v_1 < v_2$ and (10) and (11). (Notice that (11) actually rules out $p = 0$.)

In most cases the principal can do better with grade wages than with strictly monotonic wages. For example, if $\tau > 2\max\{d_\alpha, d_\beta\}$, then as we saw in Section 3, by taking $p = 1/2$ and wages $0 = v_0 < \varepsilon = v_1 < 2\varepsilon = v_2$, the principal can satisfy both of the incentive constraints (10) and (11) while paying the workers almost nothing.

Even if τ is much smaller, the principal can exploit status to reduce his wage bill. Interestingly, the possibility of combining wage incentives with status incentives leads him to create smaller elites than we saw in Section 3 with status alone. Lowering p increases the status incentive for β to work, but also lowers the status incentive for α to work. But the latter can be compensated by increasing the wages v_2 for the lucky hard worker α. Only risk aversion would prevent the principal from choosing p arbitrarily close to zero.

More precisely, suppose that $\min\{d^\alpha, d^\beta\} > \tau > 0$. Choose *any* $0 < p \le 1$. Define v_1 via equality in (10), i.e., $v_1 = d_\beta - (1-p)\tau$, and define v_2 via equality in (11), i.e., $v_2 = (d_\alpha/p) + v_1 - \tau$. Since $\tau < d_\beta$, we have $v_1 > 0$. Also, since $d_\alpha < \tau$, we have $(d_\alpha/p) - \tau > 0$ and hence $v_2 > v_1$. The grade-wage bill is

$$\begin{aligned}
(2-p)v_1 + pv_2 &= (2-p)v_1 + d_\alpha + pv_1 - p\tau \\
&= 2v_1 + d_\alpha - p\tau \\
&= 2(d_\beta - (1-p)\tau) + d_\alpha - p\tau \\
&= 2d_\beta + d_\alpha - (2-p)\tau \\
&< 2d_\beta + d_\alpha + 2\varepsilon
\end{aligned}$$

proving the superiority of grade wages over monotonic wages, even in this case. Indeed the principal will wish to choose p as close to zero as possible, making the pyramid really sharp! (Recall that (11) forbids setting $p = 0$.) But this, as we said, is an artifact of the risk-neutrality assumption. If the agents were risk-averse the optimal p would stay bounded away from zero.

Our analysis extends to many disparate types and multiple agents of each type in a straightforward manner: grade wages, by and large, are best from the principal's point of view.

The reader will have observed that we have modeled the feeling of status with a jump in utility. If we fix the rival's wage at x_m and raise n's wage x_n from below x_m, his feeling for status is reflected by a jump in utility from $-\tau$ to τ as x_n crosses x_m. This jump could be slightly smoothened and made to occur over a small interval $[x_m - \delta, x_m + \delta]$ without much affecting our analysis. (One could think of δ as a "threshold" for the feeling of status.) But what is essential is the *jump*. A *gradual* feeling of envy, given by say $\lambda(x_n - x_m)$ for $\lambda > 0$, instead of $\tau\ \mathrm{sgn}\{x_n - x_m\}$, would be helpful in reducing the wages necessary to induce hard work, but it would *not* lead to grade wages. Indeed, in our example, if we took n's utility to be $x_n + \lambda(x_n - x_m) + \tau\ \mathrm{sgn}\{x_n - x_m\}$, the superiority of grade wages would hold if, and only if, $\tau > 0$. (The incentive conditions would be $(1+\lambda)w_H^\beta \ge d_\beta$ and $(1+\lambda)(w_H^\alpha - w_L^\alpha) \ge d_\alpha$ for strictly monotonic wages; and $(1+\lambda)v_1 + (1-p)\tau \geq d_\beta$ and $(1+\lambda)p(v_2 - v_1) + p\tau \geq d_\alpha$ for grade wages, showing that the two schemes become equivalent when $\tau = 0$ even if $\lambda > 0$.)

9 References

Dubey, P. and Chien-wei Wu (2001). "When Less Scrutiny Induces More Effort," *Journal of Mathematical Economics*, 36(4), pp. 311–336.

Dubey, P. and O. Haimanko (2003). "Optimal Scrutiny in Multiperiod Promotion Tournaments," *Games and Economic Behavior*, 42(1), pp. 1–24.

Shaked, M. and J.G. Shanthikumar (1994). *Stochastic Orders and their Applications*. San Diego: Academic Press.

Spence, M. (1974). *Market Signaling: Informational Transfer in Hiring and Related Processes*. Cambridge: Harvard University Press.