GRADING EXAMS: 100, 99, 98,… OR A, B, C?

BY

PRADEEP DUBEY and JOHN GEANAKOPLOS

COWLES FOUNDATION PAPER NO. 1302

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

2010

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

Games and Economic Behavior 69 (2010) 72–94

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Grading exams: 100, 99, 98, ... or A, B, C? ☆

Pradeep Dubey [a,*], John Geanakoplos [b]

[a] Center for Game Theory in Economics, SUNY, Stony Brook and Cowles Foundation, Yale University, United States
[b] Cowles Foundation, Yale University and Santa Fe Institute, Santa Fe, NM, United States

ARTICLE INFO

Article history:
Received 2 November 2006
Available online 23 February 2010

JEL classification:
C70
I20
I30
I33

Keywords:
Status
Grading
Incentives
Education
Exams

ABSTRACT

We introduce *grading* into *games of status*. Each player chooses effort, producing a stochastic output or score. Utilities depend on the ranking of all the scores. By clustering scores into grades, the ranking is coarsened, and the incentives to work are changed.

We apply games of status to grading exams. Our main conclusion is that if students care primarily about their status (relative rank) in class, they are often best motivated to work *not* by revealing their exact numerical exam scores $(100, 99, \ldots, 1)$, but instead by clumping them into coarse categories (A, B, C).

When student abilities are *disparate*, the optimal absolute grading scheme is always coarse. Furthermore, it awards fewer A's than there are alpha-quality students, creating small elites. When students are *homogeneous*, we characterize optimal absolute grading schemes in terms of the stochastic dominance between student performances (when they shirk or work) on subintervals of scores, showing again why coarse grading may be advantageous.

In both cases, we prove that absolute grading is better than grading on a curve, provided student scores are independent.

1. Introduction

Examiners typically record scores on a precise scale $100, 99, \ldots, 1$. 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, 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 often motivates students to work harder.

☆ This is a revision of the first part of Dubey and Geanakoplos (2005), which in turn was based on Dubey and Geanakoplos (2004).

* Corresponding author.
E-mail address: pradeepkdubey@yahoo.com (P. Dubey).

0899-8256/$ – see front matter
doi:10.1016/j.geb.2010.02.001

Status is a great motivator.[1] For many people, honors conferring status, but little remuneration now or in the future, often bring forth the greatest effort.[2] 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.

To analyze the incentive effects of status, Section 2 introduces games of ordinal status, i.e., games in which the utilities are *solely* determined by the relative ranking of the players. The players choose effort levels, which then jointly yield (possibly random) scores for each player. The grading scheme converts the scores into a ranking, with ties allowed even for scores that are different. For simplicity, we focus on additive status, in which a player gains one utile for each opponent he outranks and loses one utile for each opponent who outranks him.

The designer defines a different game according to how he clumps scores into grades, coarsening the ranking. There are many possible grading schemes, and we look for those that elicit maximal effort. For concreteness we use the language of a professor grading students, though our analysis extends to any principal-agent setting.

The advantage of coarse grading can most succinctly be seen with two students α and β who have *disparate* abilities, so that α achieves a random but uniformly higher score even when he shirks and β works.[3] Suppose, for example, that β scores between 40 and 50 if he shirks, and between 50 and 60 if he works, while α scores between 70 and 80 if he shirks and uniformly between 80 and 90 if he works. With perfectly fine grading, α will come ahead of β, regardless of their effort levels. Since they care only about rank, *both* will shirk.

But, by assigning a grade A to scores above 85, B to scores between 50 and 85, and C to scores below 50, the professor can inspire β to work, for then β stands a chance to acquire the same status B as α, even when α is working. This in turn generates the competition which in fact spurs α to work, so that with luck he can get an A and distinguish himself from β. Notice that very coarse grading (giving everyone an A) would not elicit effort since then nobody has anything to gain by improving his score. Optimal grading must be coarse, but not too coarse.

Coarse grading is also useful when students are homogeneous (*ex ante* identical). For example, suppose each student scores according to the normal distribution $N(\mu, \sigma)$ with mean μ and standard deviation σ if he works, and according to $N(\hat{\mu}, \hat{\sigma})$ if he shirks, where $\mu > \hat{\mu}$ and $\sigma < \hat{\sigma}$. It is intuitively evident that an extraordinarily high score is more likely to come from a lucky shirker than from a worker. We show that the optimal grading scheme gives the same grade A to all scores above some threshold x_A, and is perfectly fine for scores less than x_A.

Coarse grading often gets students to work harder, but no doubt reduces the screening content delivered by schools. Our analysis reveals that if the schools sought to convey more information about the quality of their students, the students would work less hard and be of lower quality![4]

It should be emphasized that coarse grading does not involve what are commonly called handicaps. Handicaps discriminate between contestants by bestowing an advantage on the weak. Handicaps thus presume knowledge of individual contestants' abilities, as well as the "legality" of the discrimination. The grading we describe in this paper is, in contrast, required to be completely anonymous in that grades depend only on the exam scores of the students and not on their names. It is also required to be 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.

Let $\mathcal{S}$ denote the class of all anonymous, monotonic grading schemes. Two canonical examples, which are most often used in practice to grade exams, are absolute grading (90 to 100 gets an A, 80 to 89 gets a B and so on) and relative grading (popularly called grading on a curve: the top 5 students get an A, the next 8 get a B, and so on). Let $\mathcal{A}$ and $\mathcal{C}$ denote the classes of all absolute grading schemes and all relative grading schemes, respectively.

For any class $\mathcal{W} \subset \mathcal{S}$ we shall say that a grading scheme $\gamma \in \mathcal{W}$ is $\mathcal{W}$-optimal if no other scheme in $\mathcal{W}$ can generate higher incentives to work. In general we have found it difficult to characterize optimal grading schemes. We therefore concentrate on two special classes ($\mathcal{A}$ or $\mathcal{C}$), and on two extremal contexts (student abilities are disparate or homogeneous).

[1] Veblen (1899) famously introduced conspicuous consumption, i.e., the idea that people strive to consume more than others partly for the sake of higher status. A large empirical literature, starting from Easterlin (1974), has shown that happiness indeed depends not just on absolute, but also on *relative*, consumption.

The modeling of status has taken two forms. The cardinal approach makes utility depend on the difference between an individual's consumption and others' consumption (see, e.g., Duesenberry, 1949; Pollak, 1976; Fehr and Schmidt, 1999; Itoh, 2004; Demougin et al., 2006; Englmaier and Wambach, 2006). The ordinal approach makes utility depend on the individual's rank in the distribution of consumption (see, e.g., Frank, 1985; Robson, 1992; Direr, 2001; and Hopkins and Kornienko, 2003). Our model of status is in the ordinal tradition.

[2] This should be contrasted with the purely instrumental role status might play, for instance when higher consumption signals higher wealth and hence eligibility as a marriage partner (see e.g., Cole et al., 1992, 1995, 1998; and Corneo and Jeanne, 1997). Like the authors in the previous footnote, we take seriously the value of status to people, in and of itself.

[3] 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 analysis still applies.

[4] The "quality" of a graduating student evidently depends on both his (innate) ability and on how hard he studied. Our analysis shows that there is a trade-off between the signaling and production of quality.

In Section 3 we characterize $\mathcal{A}$-optimal grading schemes for an arbitrary number of students of disparate abilities. Our first and most important conclusion is that in order to create the largest incentives to work, the professor should always use coarse grading. Our second conclusion is that $\mathcal{A}$-optimal grading creates small elites, excluding many from membership who have equal abilities and have also worked hard but have been unlucky in the scores realized. In a population made up of equal numbers of students of three 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. If less able students have higher costs from studying hard, as Spence (1974) suggested, then the pyramiding becomes still more extreme.

In Section 4 we characterize $\mathcal{A}$-optimal grading schemes when students are homogeneous, and their score densities are independent and regular.[5] The key analytical concepts in this analysis are (first order) stochastic dominance and uniform stochastic dominance. We show that an absolute partition of scores into cells (each cell signifying a distinct grade) is $\mathcal{A}$-optimal if and only if the shirker's performance stochastically dominates the worker's inside each cell, while across cells the worker's uniformly stochastically dominates the shirker's. This enables us to construct $\mathcal{A}$-optimal grading schemes for regular score densities. We find that perfectly fine or perfectly coarse partitions are typically not $\mathcal{A}$-optimal, though we pinpoint special circumstances in which they are not merely $\mathcal{A}$-optimal but also optimal in the bigger class $\mathcal{S}$ of all anonymous grading schemes.

Given that the students only care about their relative rank, which kind of grading is better: absolute or relative ($\mathcal{A}$ or $\mathcal{C}$)? We show in Sections 3 and 5 that if the students are disparate or homogeneous, then absolute grading is always better than grading on a curve. (For instance, in the example of two disparate students α and β, grading on a curve provides no incentives whatsoever.)

The inferiority of grading on a curve is surprising, especially since it is so commonly used in practice. One explanation is that professors fear damaging their reputation if their grade profile differs too much from the school norm. Another possibility is that our theorem is no longer valid if the professor is significantly uncertain about the distribution of students' abilities, or if their scores are correlated.

Our analysis presumes for the most part that each student knows his own ability and that the students and the professor all know the *distribution* of abilities in the class. (They do not necessarily know which student has which ability.) By virtue of repeated meetings of the class, or similar classes held over many years, it is not unreasonable to suppose that this distribution can be fairly well estimated by the professor and the students alike. Nevertheless, in Section 7 we do take up the case of incomplete information with absolute grading.[6]

Our exploration of optimal grading has been carried out in a limited context: students have binary effort levels, their performances are independent, they are either disparate or *ex ante* homogeneous, and grading schemes are mostly absolute or relative (in $\mathcal{A}$ or $\mathcal{C}$). We believe some of our themes can be extended to more general settings, and we hope that this will be taken up in future work. One extension that we have undertaken is "How to pay workers when wages also confer status" (Dubey and Geanakoplos, forthcoming).

2. Games of status

In this section we precisely define what we mean by games of status, and the freedom the professor has to create grades.

Imagine a set N of students 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, as we often do, that a student's score does not depend on others' efforts; but actually several of our results do not require this independence assumption.[7] Given the exam scores $x = (x_n)_{n\in N}$, the professor must assign letter grades $\gamma(x)$. Students are assumed to care only about relative status,[8] and not about the education they are getting. We capture this by assuming 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.[9]

[5] See the density assumption in Section 4.3.

[6] Moldovanu et al. (2007) take our model and reconsider our results, replacing our hypothesis that the *distribution* of abilities in the actual class is known with the incomplete information hypothesis that student abilities are independently drawn from that distribution, so that the distribution of abilities in the actual class may be different. With a continuum of students, which we sometimes assume, the two hypotheses are the same. Moreover, with absolute grading and additive status our analysis covers the incomplete information case as well (as explained in Section 6). Only when the student population is small, and the professor grades on a curve, will there be a difference between complete and incomplete information. This is the case considered by Moldovanu, Sela and Shi.

[7] When the score x_n of one player depends (perhaps negatively) on the effort e_m, $m \neq n$ of another player, we can reinterpret our model as a parlor game.

[8] In our model a student values his grade only insofar as it defines his relative standing vis-à-vis others and places no intrinsic value on a high grade for its own sake, e.g. preferring to get a B while all others get C rather than getting an A along with everyone else. This is why we have games of "status" not of "grades."

[9] This is to keep matters simple. A "harmonic" utility might give $1/n$ utiles to a student who alone has rank n, and $(1/n + \cdots + 1/(n+m-1)) \cdot (1/m)$ utiles to each of m students who tie at rank n. Coming first instead of second provides a much bigger gain in utility than moving from 27th to 26th. Both additive and harmonic utilities are instances of "positional" status, that reward a player solely on the basis of his own position in the hierarchy.

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 scores corresponding to each grade, say $[85, 100]$ gives A, $[70, 85)$ 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.[10]

What grading scheme γ *should* a professor use, if he wants to incentivize (whenever feasible) *all* his students to put in maximal effort?[11] No matter what scheme 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 all students 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

The *strategy set* $E_n \subset \mathbb{R}_+$ of each student $n \in N$ consists of a set of effort levels that are WLOG 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 $\Delta(Y)$ denote the set of probability distributions on Y, for any set Y. The *performance map*

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

associates stochastic scores with effort levels. Here $\pi(e)$ gives the probability distribution of score vectors when the students put in effort $e \in E$.[12]

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 ranks students according to $\gamma(x)$ when the scores obtained are $x \in \mathbb{R}^N$. Each rank corresponds to a grade. Coarse grading puts different scores into the same rank. We consider, in principle, only maps γ that are anonymous and monotonic. Anonymity means that the grades depend on the scores, not on the names. Monotonicity means two things: first, if a player j scores at least as high as another player i, then j's rank is at least as high as i's; second, if j increases his score, his rank relative to each other player is at least as good as before.[13]

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 "fine." We assume throughout that any bounded interval of $\mathbb{R}$ is covered by finitely many intervals from $\mathcal{P}$. When an interval[14] $[a, b)$ is designated fine, it is taken to represent the partition $\{\{x\}\colon\ x \in [a, b)\}$ 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 $\mathcal{P}$.[15]

Fix a partition $\mathcal{P}$ as above. Then for any two scores $a, b \in \mathbb{R}$ 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 defines the absolute grading $\gamma_{\mathcal{P}} : \mathbb{R}^N \to \mathcal{R}$. 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. Random grading

We could also introduce randomness in γ without violating monotonicity or anonymity of the grading scheme. For example, the professor could announce that he will flip a coin just before grading the exam: if heads he will take the

[10] In practice students may not really know the connection between their effort and the scores they are likely to get; they may not be told explicitly how the professor plans to grade; and they may not know exactly the distribution of innate abilities in the class. The game theoretic approach nevertheless requires that students make precise conjectures about all these unknowns in order to model the competition between them as a formal non-cooperative game.

[11] We could have considered other goals, like what grading scheme would give the highest expected total score, even when it is not feasible to induce all students to exert full effort. The results would have a similar flavor, but we leave them for future research.

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

[13] That is, if $y_i = x_i$ for all $i \in N\backslash\{j\}$, and $y_j \geqslant x_j$, then (a) $j \succ_{\gamma(y)} i$ if $j \succ_{\gamma(x)} i$ and (b) $j \succcurlyeq_{\gamma(y)} i$ if $j \approx_{\gamma(x)} i$.

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

[15] Recall that students care only about their relative grade in the class. The professor could *ex ante* fix a different letter grade for each cell. Equivalently, he could wait until the realization of exam scores, and *ex post* assign the letter grade A to the highest cell that includes at least one student's score, a B to the next highest cell that includes at least one score, and so on. That way some student always gets an A, and the number of grades never exceeds the number of students in the class.

interval [86, 100] to be an A, while if tails he will count any score in the interval [84, 100] as an A. When the performance map π is deterministic, random grading may be needed to induce maximal effort.

Let us extend the grading map γ to

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

with scores being assigned random grades. We now discuss how this randomness arises naturally with grading on a curve when there are ties.

2.2.3. Grading on a curve with random tie breaking

Given distinct scores $x = (x_n)_{n\in N} \in \mathbb{R}^N$, a grading curve is defined by the vector $(n_A, n_B, \ldots)$ with $n_A + n_B + \cdots = N$. 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, and so on.

When there are ties in the scores, we break them randomly, generating many strict rankings with equal probability, and then we apply the grading curve to each of them. This generates random grades.

2.3. Utilities

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

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

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 ordinal status.

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 $\gamma : \mathbb{R}^N \to \Delta(\mathcal{R})$. Then, given effort levels $e \equiv (e_k)_{k\in N} \in E$, the *payoff* to $n \in N$ is his expected net utility = expected exam payoff – disutility of effort:

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

Here $\mathrm{Exp}_{\pi(e)}$ denotes expectation w.r.t. the distribution $\pi(e)$ over scores $x \in \mathbb{R}^N$, and $\mathrm{Exp}_{\gamma(x)}$ denotes expectation w.r.t. the distribution $\gamma(x)$ over score rankings $R \in \mathcal{R}$.[16]

2.5. Binary effort levels and incentives to work

We shall concentrate on the case of two effort levels: high (work) H_n and low (shirk) L_n, with $H_n > L_n$ for each agent n. Let $H = (H_1, \ldots, H_N)$ be the strategy profile of *maximal effort* and let $H_{-n} \equiv (H_k)_{k\in N\setminus\{n\}}$. Define the *incentive to work* created for each $n \in N$ by the grading scheme γ (assuming all others are working) to be

$$I^n(\gamma) = u^n_\gamma(H) - u^n_\gamma(H_{-n}, L_n).$$

2.6. Optimal grading

Let $\mathcal{S}$ be the class of all anonymous, monotonic grading schemes $\gamma : \mathbb{R}^N \to \Delta(\mathcal{R})$. Let $\mathcal{W} \subset \mathcal{S}$. Denote

$$I(\gamma) = \big(I^1(\gamma), \ldots, I^N(\gamma)\big),$$
$$I(\mathcal{W}) = \big\{I(\gamma)\colon\ \gamma \in \mathcal{W}\big\}.$$

We shall say that a grading scheme $\gamma \in \mathcal{W}$ is $\mathcal{W}$-*efficient* if there is no $\gamma' \in \mathcal{W}$ with $I(\gamma') \gneqq I(\gamma)$ and that it is $\mathcal{W}$-*maxmin* if there is no $\gamma' \in \mathcal{W}$ with $\min_{n\in N} I^n(\gamma') > \min_{n\in N} I^n(\gamma)$. Finally we shall say that γ is $\mathcal{W}$-*optimal* if it is both $\mathcal{W}$-efficient and $\mathcal{W}$-maxmin.

Let $d_n = H_n - L_n$ be n's disutility for switching from shirk to work and let $d = (d_1, \ldots, d_n)$. Then H is a Nash equilibrium of the game Γ_γ if and only if $I(\gamma) \geqslant d$.

Since our goal is to find grading schemes that incentivize every student to study hard, we will focus on optimal schemes.

[16] Note that it is not necessarily the case that a higher expected exam score for n means a higher exam payoff to him. For example, if grading is perfectly fine, then getting a much lower score than his rival with probability 0.49 and getting a slightly higher score with probability 0.51 yields him positive exam payoff, though he has a lower expected exam score than his rival.

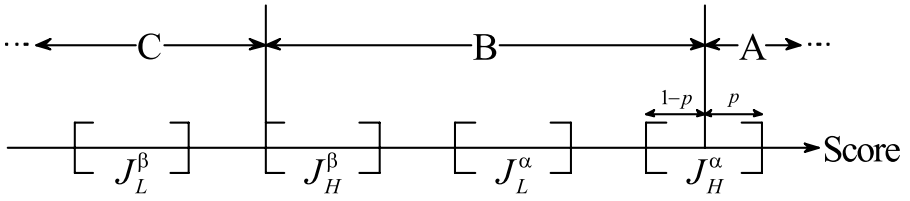

Fig. 1. The partition $\mathcal{P}(p)$.

3. Disparate students

We begin with the first of our two extremal contexts, namely the one in which students have "disparate" abilities. More precisely, suppose there are types $1, \ldots, \ell$. Each student of type i scores according to a nonatomic probability distribution with support in the interval J_H^i when he works, and any probability distribution with support on the interval J_L^i if he shirks.[17] We assume that the score of any student is independent of the effort levels and scores of the other students. This considerably simplifies our analysis, but the reader can check that many of our basic results hold even with correlation.

We call the types *disparate* if $J_L^i < J_H^i < J_L^{i+1} < J_H^{i+1}$ for every $i = 1, \ldots, \ell - 1$, where $I < J$ means that the two intervals are disjoint and I lies below J. Higher types score better than lower types no matter what effort levels either is choosing.

3.1. Coarsening

We begin with the simplest example, in which we can compute an optimal absolute grading partition, illustrating the benefits of coarse grading.

First suppose $N = \{\alpha, \beta\}$, i.e., there are just two disparate students with performance intervals given in Fig. 1 below. 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 $\Gamma_{\gamma_{\widetilde{\mathcal{P}}}}$ where $\widetilde{\mathcal{P}} \equiv \{\{x\}\colon\ x \in \mathbb{R}\}$ denotes the finest partition – even more, it is an NE in strictly dominant strategies. Neither would they work if he graded them on a curve, for then he would have to distinguish them all the time, which would be tantamount to fine absolute grading, or give them the same grade all the time, which obviously induces shirking.

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 Fig. 1. 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 $0 \leqslant p \leqslant 1$, 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). 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:

$$I^\alpha(p) = u_p^\alpha(H_\alpha, H_\beta) - u_p^\alpha(L_\alpha, H_\beta) = p - 0 = p,$$
$$I^\beta(p) = u_p^\beta(H_\alpha, H_\beta) - u_p^\beta(H_\alpha, L_\beta) = -p - (-1) = 1 - p.$$

The optimal $p^* = 1/2$ is given by

$$p^* = \arg\max_{0 \leqslant p \leqslant 1} \min\left\{I^\alpha(p),\ I^\beta(p)\right\} = 1/2.$$

Note that $I^n(1/2) = 1/2$ for both students n.

Recall that $d_n = H_n - L_n$ is n's disutility for switching from shirk to work, and that (H_α, H_β) is a Nash equilibrium if and only if $I^\alpha(p) = p \geqslant d_\alpha$ and $I^\beta(p) = 1 - p \geqslant d_\beta$. As long as $d_\alpha + d_\beta < 1$, both students can be induced to work with several different p. If $d_\alpha + d_\beta = 1$, then only $p = d_\alpha$ will do the job.

3.1.1. Multiple effort levels and less disparateness

The hypothesis of disparate students is not as strong as it seems. 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 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. It shows that any partition that induces both agents to work their hardest must pool part of J_H^α with part of J_H^β.

[17] The continuous randomness in scores if a student works is crucial to the analysis. If these scores were deterministic instead, we could still achieve the same incentives by randomizing the grading (e.g., in the example below, randomizing the cutoff to get an A).

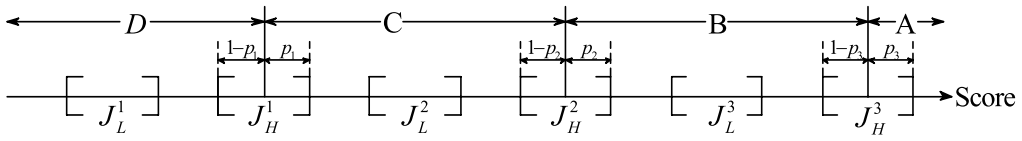

Fig. 2. The partition $\mathcal{P}(p_1, p_2, p_3)$.

3.2. Pyramiding

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

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

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

In other words, optimal grading creates a pyramid with fewer expected 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. When disutilities are severe and $d_\alpha + d_\beta = 1$, we have to pick $p = d_\alpha$ to induce work, as we just saw. This falls as d_β rises, diminishing the expected number of A's to d_α, and increasing the B's to $1 + d_\beta$.

3.2.1. Multiple students of each type

Now we show that coarsening and pyramiding persist with many students of each type. Suppose there are N_β β-type students of low ability and N_α α-type students of high ability. The reader can check that the incentive functions become:

$$I^\alpha(p) = p\delta,$$
$$I^\beta(p) = -pN_\alpha - \bigl(-(N_\beta + N_\alpha - 1)\bigr) = \bigl(1 - \mu^H p\bigr)\delta,$$

where $\delta \equiv N_\beta + N_\alpha - 1 \equiv$ utiles to a student when he beats all the others and $\mu^H \equiv N_\alpha/(N_\beta + N_\alpha - 1)$ gives the fraction of high ability in the population, when a single low-ability student stands aside. The status incentive grows *pro forma* with the population, as if there were 100 times more status in being the President of India's 1 billion than of Greece's 10 million. If this were really true it would be easier to incentive Indian students to work than Greek students. But the optimal absolute grading scheme would be virtually the same for both populations, as we see immediately below.

The optimal (maxmin) $p = 1/(1 + \mu^H)$ is obtained by solving $1 - \mu^H p = p$. When N_α and N_β are large and equal, μ^H is nearly $1/2$, and the optimal p converges to $2/3 < 1$. The pyramid remains. Indeed, the pyramid becomes more visible since the expected number of students getting a letter grade is approximately equal to the actual number of students getting that grade, by the law of large numbers.[18]

3.3. Many disparate student-types

When there are ℓ disparate 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 $0 < p_i \leqslant 1$ 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: 1 (low ability), 2 (middle ability), 3 (high ability) in Fig. 2.

Suppose there are $N_1, \ldots, N_\ell$ students of type $i = 1, \ldots, \ell$. Given the grading partition $p = (p_1, \ldots, p_\ell)$, the incentive to work for the ℓ types is

$$I^1_\ell(p) = p_1\bigl[(N_1 - 1) + (1 - p_2)N_2\bigr],$$
$$I^i_\ell(p) = p_i\bigl[(N_i - 1) + p_{i-1}N_{i-1} + (1 - p_{i+1})N_{i+1}\bigr] \quad \text{for } 2 \leqslant i \leqslant \ell - 1,$$
$$I^\ell_\ell(p) = p_\ell\bigl[(N_\ell - 1) + p_{\ell-1}N_{\ell-1}\bigr].$$

[18] Observe that 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.) This is so since $p = 1/(1 + \mu_H)$ is decreasing in μ_H. It is also interesting to observe that so long as there is at least one β student, the proportion of A's in the whole population is always less than $1/2$, since $p\mu_H = \mu_H/(1 + \mu_H) < 1/2$.

When working, a student of type $2 \leqslant i \leqslant \ell - 1$ 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 equaling) or equals (instead of losing to) every student of his own type. This gives the formula $I^i_\ell(p)$ for $2 \leqslant i \leqslant \ell - 1$. Taking $N_0 = N_{\ell+1} = 0$ in this same expression gives the formulas for $I^1_\ell(p)$ and $I^\ell_\ell(p)$.

When there are vastly more students of some types than others, an optimal partition will not necessarily equalize all the incentives. For example, suppose there are one billion students of the lowest type 1, and just two students of types 2 and 3. An efficient partition will always set $p_1 = 1$, giving an incentive to work of at least one billion (minus one) to type 1 students. A top student (of type 3) is only competing against the students of type 3 and 2, and can therefore never have incentives exceeding three utiles. This also shows that maxmin grading schemes need not be unique, since choosing $p_1 < 1$ (but not too small) will also achieve the maxmin.

Surprisingly, if $2 \leqslant N_1 \leqslant \cdots \leqslant N_\ell$, there will be a unique maxmin partition, and it will indeed equalize all the incentives, and be optimal. Furthermore, it will generate pyramiding. Indeed, each student of type $i > 1$ will have positive probability of getting a grade lower than his type.[19]

Theorem 1a. *There always exists an optimal absolute grading scheme when students are disparate.*

Theorem 1b. *Let $2 \leqslant N_1 \leqslant \cdots \leqslant N_\ell$. Then $I_\ell \equiv \max_{p\in[0,1]^\ell} \min_{1\leqslant i\leqslant \ell} I^i_\ell(p)$ is achieved at a unique $\bar{p}$; moreover, $\bar{p}_1 = 1$ and $0 < \bar{p}_i < 1$ for $i = 2, \ldots, \ell$, and all agents have the same incentive: $I^i_\ell(\bar{p}) = I_\ell \ \forall i = 1, \ldots, \ell$. Therefore $\mathcal{P}(\bar{p})$ is $\mathcal{A}$-optimal, that is, optimal in the class of all absolute grading schemes. Furthermore, there is a grading pyramid: the ratio of students obtaining the highest grade to the number of top students is equal to $\bar{p}_\ell < 1$, whereas the ratio of students getting the lowest observed grade to the number of bottom students is $\bar{p}_1 + (1 - \bar{p}_2) = 1 + (1 - \bar{p}_2) > 1$.*

Proof of Theorem 1a. The functions $I^i_\ell(p)$ are continuous on the compact set $[0,1]^\ell$. Hence $C = \arg\max_{p\in[0,1]^\ell}[\min_i I^i_\ell(p)]$ is a nonempty, compact subset of $[0,1]^\ell$. Clearly, $q \in C$ and $I(\tilde{q}) \geqslant I(q)$ together imply that $\tilde{q} \in C$. Therefore $D = \arg\max_{p\in C} \sum_{i=1}^{\ell} I^i_\ell(p)$ is also nonempty (and compact), and every point $p \in D$ is optimal. □

Proof of Theorem 1b. Since each $I^i_\ell(p)$ is continuous in p, $I_\ell(p)$ is also continuous, and so $I_\ell = \max_{p\in[0,1]^\ell} \min_{1\leqslant i\leqslant \ell} I^i_\ell(p)$ is achieved at some $\bar{p}$. Clearly any maxmin $\bar{p} \gg 0$, for otherwise $I_\ell = I_\ell(\bar{p}) = 0$, which can be bettered by choosing all $p_i = 1$.

Inspection of the formulae immediately reveals that raising $\bar{p}_i$ raises $I^i_\ell(\bar{p})$ and $I^{i+1}_\ell(\bar{p})$, but lowers $I^{i-1}_\ell(\bar{p})$. Furthermore, for any $2 \leqslant i \leqslant \ell$, if $\bar{p}_i = 1$, then from $N_1 \leqslant \cdots \leqslant N_\ell$ we get $I^i_\ell(\bar{p}) \geqslant N_i - 1 + \bar{p}_{i-1}N_{i-1} \geqslant \bar{p}_{i-1}[\bar{p}_{i-2}N_{i-2} - 1 + N_{i-1}] = \bar{p}_{i-1}[N_{i-1} - 1 + \bar{p}_{i-2}N_{i-2} + (1 - \bar{p}_i)N_i] = I^{i-1}_\ell(\bar{p})$, where the second inequality is strict if $\bar{p}_{i-1} < 1$ and $N_{i-2} \geqslant 2$.

Now we argue that for any maxmin $\bar{p}$, $I^i_\ell(\bar{p}) = I_\ell$ for all $i = 1, \ldots, \ell$. Take any maxmin $\bar{p}$ with the fewest number of coordinates i with $I^i_\ell(\bar{p}) = I_\ell$. Suppose i is the largest coordinate with $I^i_\ell(\bar{p}) = I_\ell$. If $i < \ell$, then $I^j_\ell(\bar{p}) > I^i_\ell(\bar{p})$ for all $j > i$. Lowering $\bar{p}_{i+1}$, which is possible since $\bar{p} \gg 0$, raises $I^i_\ell(\bar{p})$, and lowers the irrelevant $I^{i+1}_\ell(\bar{p})$ and $I^{i+2}_\ell(\bar{p})$. This either raises I_ℓ or reduces the number of i at which I_ℓ is attained, a contradiction either way. Hence $I^\ell_\ell(\bar{p}) = I_\ell$. Suppose $I^{i-1}_\ell(\bar{p}) > I^i_\ell(\bar{p}) = I_\ell$, for some $i = 2, \ldots, \ell$. Then from the last line of the last paragraph, $\bar{p}_i < 1$. But then raising $\bar{p}_i$ raises $I^i_\ell(\bar{p})$ and $I^{i+1}_\ell(\bar{p})$, lowering the irrelevant $I^{i-1}_\ell(\bar{p})$. This either raises I_ℓ or reduces the number of i at which I_ℓ is attained, a contradiction either way. Thus $I^i_\ell(\bar{p}) = I_\ell$ for all i and any maxmin $\bar{p}$.

Now we show that I_ℓ is achieved at a unique $\bar{p}$. Observe first that at any maxmin $\bar{p}$, $\bar{p}_1 = 1$, for if $\bar{p}_1 < 1$, increasing $\bar{p}_1$ will increase $I^1_\ell(\bar{p})$ without lowering any other $I^i_\ell(\bar{p})$, contradicting $I^1_\ell(\bar{p}) = I_\ell$ for every maxmin $\bar{p}$. But $\bar{p}_1 = 1$ and $I^1_\ell(\bar{p}) = I_\ell$ uniquely determines $\bar{p}_2$. But then $\bar{p}_1$, $\bar{p}_2$, and $I^2_\ell(\bar{p}) = I_\ell$ uniquely determines $\bar{p}_3$, and so on.

Observe that if $\bar{p}_1 = \bar{p}_2 = 1$, then obviously $I^1_\ell(\bar{p}) < I^2_\ell(\bar{p})$, contradicting all $I^i_\ell(\bar{p}) = I_\ell$. This shows $\bar{p}_2 < 1$. We showed earlier that for any $3 \leqslant i \leqslant \ell$, if $\bar{p}_{i-1} < 1$ and $\bar{p}_i = 1$, then $I^i_\ell(\bar{p}) > I^{i-1}_\ell(\bar{p})$, contradicting their equality. Thus we have shown that $\bar{p}_i < 1$ for all $i = 2, \ldots, \ell$.[20] □

[19] It is worth noting that with one student of each of three types, the optimal $(p_1, p_2, p_3) = (1, 1/2, 1)$ yielding incentive $1/2$ to each, so that the expected number of A's $= 1$, of B's $= 1/2$ and of C's $= 3/2$, giving us pyramiding but not in the strongest sense. But even here, if we introduce the Spence condition $d_3 \ll d_2 \ll d_1$ on disutility of effort, the inequalities $I^i(p) \geqslant d_i$ will (as is obvious) require $p_3 < p_2 < p_1$ by way of a solution, bringing back the full pyramid.

[20] One more observation. Consider an infinite sequence of disparate types with populations $1 \leqslant N_1 \leqslant N_2 \leqslant \cdots$. For each ℓ, let I_ℓ be the maxmin incentive for the status game with types $1, \ldots, \ell$, as above. Then I_ℓ is monotonically increasing in ℓ, converging to some $I^* \leqslant N_1 + N_2 - 1$ as $\ell \to \infty$.

To verify this, let $I_\ell = I_\ell(\bar{p})$. Define $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_\ell, \hat{p}_{\ell+1}) = (\bar{p}_1, \ldots, \bar{p}_\ell, 1)$. Then $I_{\ell+1} \geqslant I_{\ell+1}(\hat{p})$. But $I^i_{\ell+1}(\hat{p}) = I^i_\ell(\bar{p}) = I_\ell$ for all $i = 1, \ldots, \ell$. Moreover, $I^{\ell+1}_{\ell+1}(\hat{p}) = \hat{p}_{\ell+1}((N_{\ell+1} - 1) + \hat{p}_\ell N_\ell) = 1 \cdot ((N_{\ell+1} - 1) + \bar{p}_\ell N_\ell) \geqslant \bar{p}_\ell(N_\ell - 1) + \bar{p}_{\ell-1}\bar{p}_\ell N_{\ell-1} = \bar{p}_\ell(N_\ell - 1 + \bar{p}_{\ell-1}N_{\ell-1}) = I_\ell$. But $I_\ell \leqslant I^1_\ell \leqslant N_1 + N_2 - 1$ for all ℓ.

3.3.1. Work as the unique Nash equilibrium

We have just shown how to construct a partition for an absolute grading scheme that uniformly gives each student the maximum incentive to work. This partition has the property that it never cuts the shirk performance interval of any type. We now show that whenever all work is a (strict) Nash equilibrium for a partition with this property, it is the unique Nash equilibrium.

To prove this, consider a student i of the lowest type. We shall show that his incentive to work against j remains at least as high if j shirks. If j is the same type as i, then this incentive is in fact unchanged by symmetry:

$$u(H,H) - u(L,H) = u(H,L) - u(L,L)$$

because

$$u(H,H) = 0 = u(L,L),$$

$$u(H,L) = -u(L,H).$$

Now let $j > i$ be any type above i. Clearly

$$u(H,L) \geqslant u(H,H)$$

by monotonicity of performance and of the grading scheme. Note also that the partition must cut the work interval of i, or be between the shirk and work interval of i. Otherwise, given the property we have assumed, i would have no incentive to work. But this means i comes below j for sure when i shirks, no matter what j does. Hence,

$$u(L,L) = u(L,H)$$

and so finally

$$u(H,L) - u(L,L) \geqslant u(H,H) - u(L,H).$$

Thus the bottom type workers will work hard in every Nash equilibrium.

We conclude the proof by induction. Consider a worker i of any type, and assume all lower types are working. The incentive for him to work against these lower types is then exactly as in the all work Nash equilibrium. But as we just argued, the incentive for him to work remains the same against his own type, and remains the same or goes up against higher types. This concludes the proof that whenever all work is a strict Nash equilibrium, it is the unique Nash equilibrium. (And, clearly, when all work is Nash, it is a strict Nash for almost all disutilities of effort.)

3.4. 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. When all the students are disparate (or homogeneous – see Section 5), it is always better to grade according to an absolute scale, no matter how many students are in the class.[21] The reason is intuitively as follows. By moving cut points p_i on each type's work performance interval (see Fig. 2), we can continuously vary the expected number of students in any absolute letter grade. Thus, starting at the top grade and descending down the grade ladder, it is possible to reproduce any grading on a curve via an absolute scheme (as shown precisely in the proof below). But then optimal absolute grading cannot be surpassed by any grading on a curve.

Theorem 2 *(Absolute grading beats grading on a curve). Consider multiple disparate students, as in Section* 3.3. *For any grading on a curve, there exists an absolute grading scheme that generates at least as much incentive for every student as the curve. Consequently, every optimal absolute grading scheme generates higher minimum incentive than any grading on a curve.*

Proof. Let there be N_i students of type $i = 1, \ldots, \ell$, as in Section 3.3. Grading on a curve means specifying integers $K = (K_A, K_B, \ldots, K_Z)$ with $K_A + K_B + \cdots + K_Z = N = N_\ell + \cdots + N_1$, where the top K_A student exam scores get A, the next K_B get B and so on. (The probability of ties is zero.)

If there is only one disparate student of each type, then the student of type i will score below $\ell - i$ students and above $i - 1$ students whether he works or shirks. With grading on a curve, his letter grade must therefore be independent of his effort, and so grading on a curve provides no work incentive whatsoever.

Consider a general population $N = (N_1, \ldots, N_\ell)$, and any grading on a curve $K = (K_A, K_B, \ldots, K_Z)$. We can find an absolute grading scheme that creates the same incentives to work for types $2, \ldots, \ell$, and at least as much for type 1.

Let μ_i measure the distribution of scores of a type i agent when he works (so $\mu_i(T) = \text{Prob}(x_i^H \in T)$ for every $T \subset \mathbb{R}$). Define $\mu \equiv \sum_{i=1}^{\ell} N_i \mu_i$. For any relative grade G, cut $\mathbb{R}$ at the *minimum* point x such that

$$K_A + \cdots + K_G = \mu[x, \infty).$$

It is easy to check that the absolute partition defined by these cuts does the job. □

[21] Of course, with a continuum of students of each type, there is no difference between grading on a curve and absolute grading. Giving an A to the top 10% of students can be replicated by giving an A to all scores above x_A, for some appropriate threshold x_A.

3.5. More general grading schemes

Absolute grading, though better than grading on a curve, is not optimal in the class of all anonymous and monotonic grading schemes.

Recall the case of N_β β-type students of low ability and N_α α-type students of high ability, and the $\mathcal{A}$-optimal grading scheme described in Section 3.2.1. We shall now present a grading scheme that does better.

When every agent of α-type is working hard, so that no score lies in J_L^α, we assign the same grades as in our $\mathcal{A}$-optimal scheme (with $p_\beta = 1$ and $p_\alpha = 1/(1 + N_\alpha/(N_\beta + N_\alpha - 1))$). However, if there is any score in J_L^α, then apply the absolute grading scheme, with $p_\beta = p_\alpha = 1$. This new grading scheme is the same as before, except that when some α-type student shirks, more As are expected to be given (if $N_\alpha - 1 > 0$), though never to the shirker. Thus if $N_\alpha > 1$, the payoff to the shirker is lower than before. Hence the incentive of the α-type goes up, while the incentive of the β-type is unchanged.[22]

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, i.e. the same map from effort to random scores on the exam. They are free to choose different effort levels, and their scores are subject to random shocks that may give them different scores even when they choose the same effort levels. We show that coarsening still has a role to play.

Homogeneity simplifies our task in several ways. First, the incentives of all the players are aligned. Maxmin and optimal become identical. Second, when all the students work, each has an expected exam payoff of zero. This is so because the sum of their exam payoffs is zero for every *ex post* realization of scores, and because they are *ex ante* identical. Hence the incentive to work is simply the negative of the expected exam payoff to a sole shirker when the remaining $N - 1$ students are working.

It follows that there is *no* simplification gained by assuming that each student's performance is independent of the others' effort levels. For example, if we let (f, f, g) be the score densities of the (worker, worker, shirker), then we could compute directly the expected payoff of the shirker. If the shirker switched to working and we had assumed independence, we would then know that the densities would become (f, f, f). Without independence we might have to deal with new densities (h, h, h). But this does not alter the computations, since the exam payoffs with densities (h, h, h) are zero, just as they would be with (f, f, f); h would be irrelevant.

The following conditional independence simplifies the analysis as much as would have been achieved by assuming full independence.

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

We shall maintain this assumption for the rest of the paper. It immediately implies that the incentive to work is the probability that a shirker is ranked lower than a worker minus the probability that his score is ranked above the worker's, all multiplied by $N - 1$.

After presenting several examples, we give sufficient conditions for perfectly fine or perfectly coarse grading to be optimal in the class $\mathcal{S}$ of all anonymous, monotonic schemes. In general we are unable to pin down the optimal grading scheme in $\mathcal{S}$. However, under weak regularity assumptions on the score densities, we are able to characterize optimal schemes in the class $\mathcal{A} \subset \mathcal{S}$ of absolute grading schemes. In Section 5 we shall prove that absolute grading gives better incentives than grading on a curve.

4.1. Examples

We present four examples that can be encompassed in our theory. Example 1 illustrates the advantage of coarse grading when score densities are piecewise differentiable. Example 2 shows that in some circumstances perfectly fine grading is optimal. Example 3 illustrates that our theory is applicable with discrete distributions, and that even here coarse grading has a role to play. Example 4 shows that the theory can be used for the important case of normal distributions.

Consider a 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 $f(x) = 2$ if $50\% \leqslant x \leqslant 100\%$, and 0 otherwise, independent of the others' scores and effort levels. If he shirks, suppose his score has density $g(x) = 2x$ for $0 \leqslant x \leqslant 100\%$, and 0 otherwise, again independent of the others (see Fig. 3).

The probability the shirker comes behind a worker is $\int_0^{1/2} 2x\,dx + \int_{1/2}^1 2x2(1-x)\,dx = 7/12$. The probability the shirker comes ahead of a worker is therefore $1 - 7/12 = 5/12$, and we conclude that, with perfectly fine grading, shirking gives an

[22] There are natural ways of cutting down the class $\mathcal{S}$ of grading schemes to $\mathcal{S}^*$, so that any $\mathcal{A}$-optimal scheme is also $\mathcal{S}^*$-optimal. For instance, consider the following "independence of irrelevant score changes" hypothesis (satisfied by absolute grading and grading on a curve): if a student i's score falls to a level still higher than the score of a student j, then j's payoff is unchanged relative to any student $k \neq i$ whose score is above his. The proof of Theorem 1 in Dubey and Geanakoplos (2005) confirms all this when there are two disparate types.

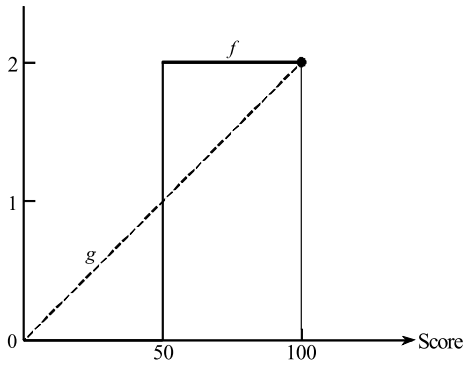

Fig. 3. Score densities.

expected exam payoff $(N-1)(\frac{5}{12}-\frac{7}{12})=-\frac{1}{6}(N-1)$. 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 $-1\int_0^{1/2}2x\,dx+0\int_{1/2}^1 2x\,dx=-1/4$ 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.

We will show later that our partition of scores

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

into just two grades yields the optimal absolute grading partition.

With three students the incentive to work under $\mathcal{P}$ is $(1/4)(3-1)=1/2$. 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 $1/3$). 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 $63/324<1/2$. Once again optimal absolute grading is better than any grading on a curve.

Next we give an example where fine grading is optimal. Suppose an exam contains K questions, and that a student who studies has (independent) probability p of getting each question right, while if he shirks the probability drops to $q<p$. Theorem 4 will imply that for this example, perfectly fine absolute grading is optimal in the class $\mathcal{S}$ of all anonymous and monotonic schemes.

Our third example shows that coarseness can be important even when scores are discrete. Imagine an exam with two questions covering the two halves of the course. Suppose that if a student studies hard, he has probability $p=0.6$ of getting any question right, independently across questions. If he shirks and studies only half the course, he has probability $q=0.8$ of getting the corresponding question right, and zero chance of getting the other question. Thus the probabilities for getting $(0,1,2)$ questions right are $(0.2,0.8,0)$ for the shirker and $(0.16,0.48,0.36)$ for the worker. Clearly the hard working student will do better most of the time (his score stochastically dominates the shirker's score). How should the professor grade the exams?[23]

Suppose there are just two students, and that the second student studies hard. What incentive does the first student have to work? Fine grading gives the shirker an expected exam payoff of $(1-p)^2q-[p^2q+p^2(1-q)+2p(1-p)(1-q)]=-0.328$. If both students work, then by symmetry and the fact that total exam payoff is inevitably zero, the expected utility of each is 0. The incentive to work with fine grading is thus 0.328.

Suppose instead that the professor uses just two grades, an A for a perfect exam, and a B for anything else. Then the expected exam payoff of a shirker is $-[p^2q+p^2(1-q)]=-p^2=-0.36$. His incentive to work is thus 0.36, since again if they both work, each has an expected exam payoff of zero. Since $0.36>0.328$, we see that coarse grading gives higher incentives to work.

As a final example, suppose that we are grading the relative performances of two hedge funds. Suppose that a hedge fund that works on research will generate log returns normally distributed with mean μ and standard deviation σ, while a hedge fund that shirks will generate returns distributed according to $\hat{\mu}<\mu$ and $\hat{\sigma}>\sigma$. If the fund managers cared only about their relative grade, would they be motivated to work harder if the grade was simply their return? Intuitively that seems wrong, since an extraordinarily high return is more likely from the high variance shirker, despite his lower expected return. We shall see that the optimal grading scheme indeed does not reward higher returns after some point.

[23] Suppose the shirker could study either half of the course, so the professor cannot distinguish the two students by attaching higher weight to the second question. We assume his grading depends only on the total number of correct answers of each student.

4.2. The general theory with iid students

We turn first to the general situation, identifying conditions under which it is optimal, in the class of *all* anonymous monotonic grading schemes, to completely reveal the scores or to completely mask them.

A key role in all that follows is played by the notion of stochastic dominance.

4.2.1. Stochastic dominance

Definition. When we say "conditional on x and y being in $[a,b)$ something happens" we mean that it happens if $P(x \in [a,b) \wedge y \in [a,b)) > 0$.

Definition. We say that the *random variable* x (*stochastically*) *dominates the independent random variable* y on the interval $[a,b)$ if, conditional on both being in $[a,b)$, x (first order) stochastically dominates y, i.e.

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

or,

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

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

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

If the inequality is strict for all $\theta \in (a,b)$, we write $x \succ y$ on $[a,b)$ and call it strict dominance. If $[a,b) = (-\infty,\infty)$, then we simply write $x \gtrsim y$ or $x \succ y$.

Stochastic dominance has an extremely important role to play in monotonic grading schemes, including absolute grading.

Lemma 1. *Suppose $x \gtrsim y$. Let the exam scores x and y be independent of the exam scores of every student $n = 1,\ldots,N-1$. Let γ be any monotonic grading scheme for N students. Then the expected exam payoff to the last student 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 $\hat{x} \geqslant \hat{y}$ 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 follows that

Theorem 3 *(When perfectly coarse grading is optimal). If a shirker has exam scores distributed according to x_L while the worker's is distributed according to x_H, and $x_L \gtrsim x_H$, then no anonymous, monotonic grading scheme can provide any incentive to work. We might as well give all students an A.*

In the first three examples of this section, the worker scores stochastically dominated the shirker's scores, and indeed this is what gave the student an incentive to work under all the grading schemes.

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 $x \gtrsim_U y$ on $[A,B)$.

Uniform domination can be characterized in terms of likelihood ratios in a manner that makes it much more handy to work with.

Lemma 2. *Let x and y be independent on $[A,B)$ with density functions f and g, respectively. Then x uniformly dominates y on $[A,B)$ if and only if the likelihood ratio $f(t)/g(t)$ is increasing almost everywhere on $[A,B)$, where $f(t)/g(t)$ can be defined suitably arbitrarily if $f(t) = g(t) = 0$.*

Proof. This follows from Theorem 1.C.2 in Shaked and Shanthikumar (1994). □

It is critical in understanding the first and third examples of this section to observe that although the worker's scores (stochastically) dominates the shirker's, there are subintervals on which the shirker's score uniformly (stochastically) dominates the worker's. In the first example of this section x_L uniformly dominates x_H on $[50,100]$. In the third example, $q/(2p(1-p)) = 0.8/0.48 > 0.2/0.16 = (1-q)/(1-p)^2$, so on the cell $\{0,1\}$ the shirker uniformly dominates the worker.

Another instance of uniform domination occurs in the second example, where 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

$$\frac{\binom{K}{k}p^k(1-p)^{K-k}}{\binom{K}{k-1}p^{k-1}(1-p)^{K-k+1}} > \frac{\binom{K}{k}q^k(1-q)^{K-k}}{\binom{K}{k-1}q^{k-1}(1-q)^{K-k+1}}$$

or

$$\frac{p}{1-p} > \frac{q}{1-q}.$$

Thus if $p > q$, the first score uniformly dominates the second score over the range of all scores.[24]

When f and g are differentiable, $f(t)/g(t)$ is increasing if and only if $f'(t)/f(t) \geqslant g'(t)/g(t)$. 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

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

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. This is crucial in understanding the fourth example.[25]

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 has score x_L with density g while each other student k has an independent score $x_k \sim x_H$, with density f. (Here $\sim$ denotes identical in distribution.) We begin with a simple theorem showing that when $x_H \gtrsim_U x_L$, perfectly fine grading is optimal, even in the wider class $\mathcal{S}$ of all monotonic and anonymous grading schemes.

Theorem 4 *(When perfectly fine grading is optimal). Let there be N iid students who can work or shirk. Suppose $x_H \gtrsim_U x_L$. Further, suppose x_H and x_L are either discrete or have piecewise continuous densities f and g with no atoms. Then perfectly fine grading is optimal in the class $\mathcal{S}$ of all monotonic, anonymous grading schemes.*

Proof. Consider the case of two of the students, the shirker L and one of the workers H. Fix the scores of all the other workers. A monotonic grading scheme γ gives a different payoff (from fine) to L against H precisely on the set

$$W = \big\{(x_L,x_H) \in \mathbb{R}^2\colon\ x_L \neq x_H,\ \text{yet } \gamma \text{ gives } x_L \text{ and } x_H \text{ the same grade}\big\}.$$

By anonymity, $(\alpha,\beta) \in W$ if and only if $(\beta,\alpha) \in W$. Suppose first that the densities f and g are discrete. By independence, f and g also give the densities of scores for H and L conditional on the other scores being fixed. From the uniform domination $x_H \gtrsim_U x_L$, we know that if $\beta > \alpha$, then $f(\beta)/g(\beta) \geqslant f(\alpha)/g(\alpha)$. Hence replacing the masking grading γ on W with fine grading *lowers* the expected exam payoff to the shirker by

$$\sum_{\substack{(\alpha,\beta)\in W\\ \alpha<\beta}} \big[f(\beta)g(\alpha) - g(\beta)f(\alpha)\big] \geqslant 0.$$

Next suppose that f and g are piecewise continuous. Since W is measurable it can be approximated arbitrarily closely (in Lebesgue measure) by a union of small rectangles $Q_{\alpha,\beta} = \{(x_L,x_H)\colon\ \alpha-\varepsilon \leqslant x_L < \alpha+\varepsilon \text{ and } \beta-\eta \leqslant x_H < \beta+\eta\}$ whose interiors do not contain any points of discontinuity of f and g. By anonymity, we may assume that the mirror square $Q^* = \{(x_L,x_H)\colon\ \beta-\eta \leqslant x_L < \beta+\eta \text{ and } \alpha-\varepsilon \leqslant x_H < \alpha+\varepsilon\}$ is also part of the approximation. These squares have area approximately equal to $4\varepsilon\eta f(\alpha)g(\beta)$ or $4\varepsilon\eta g(\alpha)f(\beta)$. If $\beta > \alpha$, then by uniform stochastic dominance, $f(\beta)g(\alpha) \geqslant$

[24] 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.

[25] The attentive reader might be puzzled, since the binomial exam scores "converge" to normal as $K \to \infty$, yet we never see the tail where x_q dominates x_p. That is because this tail is always beyond K. In fact it is not the exam scores, but normalized exam scores, which converge to normal, and the means of x_p and x_q are diverging at the rate K (not $\sqrt{K}$).

$g(\beta)f(\alpha)$. The masking on W thus hides the fact that x_H would have come ahead of x_L more often than behind x_L when both variables are in W. Hence masking W does not improve the incentive to work. Integrating over all possible fixed scores of the other workers shows that the expected payoff of L against H is lower when grading is perfectly fine. □

Domination and uniform domination can be defined exactly the same way for any totally ordered set, such as a partition $\mathcal{P}$. The likelihood ratio criterion for uniform domination appearing in Lemma 2 also carries over to partitions. Given a density f and a partition $\mathcal{P}$, define the density

$$f_{\mathcal{P}}(x) = \begin{cases} f(x) & \text{if } x \in [a,b) \text{ a perfectly fine cell in } \mathcal{P}, \\ \frac{1}{b-a}\int_a^b f(t)\,dt & \text{if } x \in [a,b) \text{ a masked cell in } \mathcal{P}. \end{cases}$$

The analogue of Lemma 2 still holds: x uniformly dominates y on $[A,B)$ with respect to $\mathcal{P}$ if and only if $f_{\mathcal{P}}(t)/g_{\mathcal{P}}(t)$ is increasing on $[A,B)$. More importantly, Theorems 3 and 4 and their proofs all hold for totally ordered sets consisting of cells of any arbitrary absolute partition.[26] In particular, if x_H uniformly dominates x_L on $(-\infty,\infty)$ with respect to $\mathcal{P}$, then grading according to $\mathcal{P}$ penalizes a shirker at least as much as any anonymous, monotonic grading scheme that cannot distinguish scores that are indistinguishable in $\mathcal{P}$.

We have found conditions for perfectly coarse and perfectly fine grading to be optimal in $\mathcal{S}$. The general characterization of optimal schemes in $\mathcal{S}$ is quite elusive. So we turn to the narrower class $\mathcal{A}$ of absolute grading schemes, where we can completely characterize optimal grading. Later we show that an optimal absolute grading partition beats any grading on a curve.

4.3. Optimal absolute grading partitions: $\mathcal{A}$-optimality

Our characterization of optimal absolute grading partitions begins by asking whether the shirker's payoff can be lowered below zero by a single cut at θ, creating a two cell partition $\{(-\infty,\theta),[\theta,\infty)\}$. We shall find that if the worker's score distribution f stochastically dominates the shirker's score distribution g, then any cut will help, while under the reverse domination, every cut will hurt.

Next we consider a partition and ask whether cutting a cell in the partition into two cells will further help incentives or set them back. The answer depends on who is the better player, *conditional* on both scores lying inside the same cell. If the shirker is better, we must not reveal this, since we are trying to minimize his score, and keep the cell uncut. For example, the shirker may be very unlikely to get a score above 90. But conditional on both the shirker and worker getting above 90, it may be more likely that the shirker does better. (The shirker may have memorized the answers to last year's exam. In the unlikely event that this year's exam questions are the same he will get 100; otherwise he will get 0.)

The guiding principle in creating optimal partitions is to mask regions of the score space where the shirker is better than the worker, and to ensure that across cells the worker is better, so that the partition reveals the deficiencies of the shirker. This characterization will be used in Section 5 to prove that absolute grading is better than grading on a curve.

Lemma 3. *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*

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

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

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

With independence, this becomes

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

The second case is analogous. □

[26] Theorem 3 in fact holds for an arbitrary ordered set by the same proof. For Theorem 4 we must adjust the proof for the case of orders derived from partitions. In the second half of the proof of Theorem 4, allow rectangles $Q_{\alpha,\beta}$ where $\alpha \in [a,b)$ a masked cell in $\mathcal{P}$, only if $\alpha = (a+b)/2$ and $\varepsilon = (b-a)/2$. Similarly, if $\beta \in [a,b)$, a masked cell in $\mathcal{P}$, then only consider the rectangle $Q_{\alpha,\beta}$ if $\beta = (a+b)/2$ and $\eta = (b-a)/2$.

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*

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

and thus increases if and only if

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

Proof. The first display 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 $\mathcal{P}^*$. The first inequality of the second display just rearranges terms, and the second inequality follows by noting that if x_L is relatively more likely (than x_H) to fall in the right half of the interval, then it is less likely to fall in the left half. □

We are now ready to state some theorems about the optimal absolute grading partition. It turns out that stochastic dominance plays the central role in determining whether or not there should be masking (i.e., giving the same grade to different scores).

Theorem 5 *(Coarseness in the optimal absolute grading). Let there be N iid students who can work or shirk. Suppose that on some interval $[a,b)$, x_L 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 $[a,b)$. Furthermore $\mathcal{P}^*$ adds no extra cuts to $\mathcal{P}$ except possibly at a or b.*

If x_L strictly dominates x_H on $[a,b)$, then every optimal grading partition is coarse on $[a,b)$.

Proof. Consider the following picture:

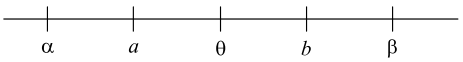

Fig. 4. Cutting (a,b) at θ.

Let $\mathcal{P}$ be a partition consisting entirely of perfectly masked cells. Let α be the cut at the highest score less than or equal to a, and let β be the cut at the lowest score greater than or equal to β, where $-\alpha$ or β might be infinite. Then we may write $\mathcal{P} = \{[\alpha,\theta_1),\ldots,[\theta_k,\beta)\}$, where $a < \theta_1 < \cdots < \theta_k < b$. We shall show that it is always possible to find another partition with no cuts inside $[a,b)$ without reducing incentives to work. 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, writing (e.g.) $P_L[c,d) \equiv P(x_L \in [c,d))/P(x_L \in [\alpha,\beta))$ for any $[c,d) \subset [\alpha,\beta)$.

If adding a cut at a increases incentives, then add it and WLOG take $a = \alpha$. Otherwise by Corollary 3 we must have

$$\frac{P_L[\alpha,a)}{P_H[\alpha,a)} \leqslant \frac{P_L[a,\theta_1)}{P_H[a,\theta_1)}.$$

Since x_L dominates x_H on $[a,b)$, we know that

$$\frac{P_L[a,\theta_1)}{P_H[a,\theta_1)} \leqslant \frac{P_L[\theta_1,b)}{P_H[\theta_1,b)}.$$

It follows that

$$\frac{P_L[\alpha,a)}{P_H[\alpha,a)} \leqslant \frac{P_L[a,\theta_1)+P_L[\theta_1,b)}{P_H[a,\theta_1)+P_H[\theta_1,b)} = \frac{P_L[a,b)}{P_H[a,b)}$$

and therefore

$$\frac{P_L[\alpha,b)}{P_H[\alpha,b)} = \frac{P_L[\alpha,a)+P_L[a,b)}{P_H[\alpha,a)+P_H[a,b)} \leqslant \frac{P_L[a,b)}{P_H[a,b)}.$$

Since we will not contemplate any cuts between α and a, we shall treat the cell $[\alpha,a)$ as a single element We shall now demonstrate that x_L dominates x_H on the ordered set $\{[\alpha,a)\} \cup [a,b)$, where $[a,b)$ represents the usual continuum of elements. Take any cut $c \in [a,b)$. Then it follows from the last display that

$$\frac{P_L[c,b)/P_L[\alpha,b)}{P_H[c,b)/P_H[\alpha,b)} \geqslant \frac{P_L[c,b)/P_L[a,b)}{P_H[c,b)/P_H[a,b)}.$$

Since x_L dominates x_H on $[a,b)$, we know that

$$\frac{P_L[c,b)/P_L[a,b)}{P_H[c,b)/P_H[a,b)} \geqslant 1.$$

Since c was arbitrary, we conclude that x_L dominates x_H on $[\alpha,b)$.

By an identical argument we can now show that either $b=\beta$, or x_L dominates x_H on the ordered set $\{[\alpha,a)\}\cup[a,b)\cup\{[b,\beta)\}$.

By Theorem 4 applied to ordered sets, we conclude that removing all the cuts in $[a,b)$ from the partition $\mathcal{P}$ will not reduce the incentive to work. □

Theorem 5 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 any $\mathcal{A}$-optimal partition.

Theorem 5 also leads to necessary and sufficient conditions for the $\mathcal{A}$-optimality of a partition $\mathcal{P}$. We say that x_L dominates x_H *inside* a partition $\mathcal{P}$ if $x_L \gtrsim x_H$ on $[a,b)$ for every cell $[a,b)\in\mathcal{P}$. We say that x_H uniformly dominates x_L *outside* a partition $\mathcal{P}$ if $x_H \gtrsim_U x_L$ on $\mathcal{P}$, i.e., across the ordered set of cells of $\mathcal{P}$.

Theorem 6 *(Inside domination and uniform outside domination imply optimality). Let there be N iid students who can work or shirk. A partition $\mathcal{P}$ is an optimal absolute grading partition if x_L dominates x_H inside $\mathcal{P}$ and x_H uniformly dominates x_L outside $\mathcal{P}$.*

Proof. Take a bounded interval $I\subset(-\infty,\infty)$ such that the probability that both x_L and x_H are in I is at least $1-\varepsilon$. Consider a finite cover of I by consecutive intervals in $\mathcal{P}$ whose union $[a,b)\supset I$. Let $I_1=[a_1,b_1), I_2=[a_2,b_2),\ldots,I_k=[a_kb_k)$ be all the intervals in the cover that are not perfectly fine.

Suppose $\mathcal{P}'$ is any partition. If $\mathcal{P}'$ cuts interval I_1, by Theorem 5 we can remove those cuts, possibly adding new cuts at a_1,b_1, obtaining a new partition $\mathcal{P}'_1$ which does at least as well as $\mathcal{P}'$. Note that the new cuts, if any, do not cut any of the intervals $I_1,I_2,\ldots,I_k$. In general, given $\mathcal{P}'_i$, use the process defined by the proof of Theorem 5 to remove the cuts of $\mathcal{P}'_i$ that lie inside $[a_{i+1},b_{i+1})$, possibly adding new cuts at a_{i+1},b_{i+1}, to obtain a partition $\mathcal{P}'_{i+1}$ that does at least as well as $\mathcal{P}'_i$. Iterate the process k times to arrive at $\mathcal{P}'_k$. $\mathcal{P}'_k$ does at least as well as $\mathcal{P}'$, and does not cut any cell of $\mathcal{P}$ lying in $[a,b)$. Thus restricted to $[a,b)$, $\mathcal{P}'_k$ is a coarsening of $\mathcal{P}$. Since $x_H\gtrsim_U x_L$ across the cells of $\mathcal{P}$ it follows from Theorem 4 that, conditional on both x_L and x_H lying in $[a,b)$, the payoff to x_L is (weakly) lower in $\mathcal{P}$ than in $\mathcal{P}'_k$. But the difference in incentive between $\mathcal{P}'_k$ and $\mathcal{P}'$ is at most $(N-1)\varepsilon$. Since ε was arbitrary, $\mathcal{P}'$ cannot provide strictly better incentive than $\mathcal{P}$. □

Suppose that all students are homogeneous, with independent, and normally distributed exam scores. If work raises a student's expected exam score, without changing its variance, then Theorem 5 implies that an $\mathcal{A}$-optimal grading scheme is perfectly fine.

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 $\mathcal{A}$-optimal grading scheme is to reveal the exact scores.

But consider the first and third examples 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 6 guarantee that these are indeed optimal partitions. In the third example, x_L uniformly dominates x_H on $\{0,1\}$, while x_H uniformly dominates x_L across the partition cells $\{0,1\}$, $\{2\}$, since $0.36/0.64>0/1$. In the first example, 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 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 $f_{\mathcal{P}}(t)=0$, and that of a shirker as $g_{\mathcal{P}}(t)=0.5$. On $[50,100)$ the effective densities become $f_{\mathcal{P}}(t)=2$ and $g_{\mathcal{P}}(t)=1.5$. Clearly $f_{\mathcal{P}}(t)/g_{\mathcal{P}}(t)$ is increasing.

In our next theorem we show that inside domination and uniform outside domination are also *necessary* conditions for a partition to be $\mathcal{A}$-optimal, when agents are homogeneous. For the theorem we need to impose slightly stronger conditions.

Density assumption. We assume that one of the following three conditions holds: (1) x_H and x_L all both discrete; (2) x_H and x_L both have piecewise continuously differentiable densities f and g, and furthermore, at any (isolated) point x of discontinuity, $f_+(x)$, $f_-(x)$, $g_+(x)$, $g_-(x)$ all exist; (3) x_H and x_L, with densities f and g, are *generic*, i.e. there is a countable set $\{\cdots<a_i<a_{i+1}<\cdots\}$ such that $f(t)$ and $g(t)$ are continuous and $f(t)/g(t)$ is continuous and strictly increasing on $[a_i,a_{i+1})$ for all even i and strictly decreasing for all odd i.

The density assumption allows us to clarify the roles of uniform domination and domination.

Lemma 4 *(Uniform domination vs. strict domination). Let $[a,b)$ be any (possibly infinite) interval, and let the random variables x_H and x_L satisfy the density assumption. Then either $x_H\gtrsim_U x_L$ on $[a,b)$, or else there is a subinterval $[c,d)\subset[a,b)$ on which $x_L\succ x_H$.*

Proof. First let the random variables be discrete. Either $f(t)/g(t)$ is weakly increasing everywhere on $[a,b)$, or else there is a subinterval $[c,d)$ on which $f(t)/g(t)$ is strictly decreasing. (The subinterval $[c,d)$ may contain only two consecutive points

in the support of either f or g.) By Lemma 2, we get the even stronger conclusion that either $x_H \gtrsim_U x_L$ on $[a,b)$, or else $x_L \succ_U x_H$ on $[c,d)$.

If the random variables are generic, as in case 3, we get the same conclusion by exactly the same argument.

Now consider case 2. If the derivative of the function $f(t)/g(t)$ is negative at some interior point $a < x < b$, then there is a small interval $[c,d)$ containing x on which $f(t)/g(t)$ is strictly decreasing. By Lemma 2, $x_L \succ_U x_H$ on $[c,d)$.

If $f(t)/g(t)$ jumps down at some non-differentiable point $a < x < b$, then by hypothesis $f(t)/g(t)$ is differentiable at all $t \in [c,x) \cup (x,d)$ for any small enough interval $[c,d) = [x-\varepsilon, x+\varepsilon)$ containing x. Now let $\theta = x - \delta$ be any cut of $[c,d)$ to the left of x. (Cuts to the right of x can be handled exactly the same way.) Then $\delta < \varepsilon$. By continuity, for very small ε, we must have

$$\frac{P[x_L \in [\theta,d) | x_L \in [c,d)]}{P[x_L \in [c,d)]} \Big/ \frac{P[x_H \in [\theta,d) | x_H \in [c,d)]}{P[x_H \in [c,d)]} \approx \frac{g(x_-)\delta + g(x_+)\varepsilon}{g(x_-)\varepsilon + g(x_+)\varepsilon} \Big/ \frac{f(x_-)\delta + f(x_+)\varepsilon}{f(x_-)\varepsilon + f(x_+)\varepsilon} > 1,$$

where the last inequality follows from $\delta/\varepsilon < 1$ and the fact that

$$\frac{g(x_-)}{g(x_+)} < \frac{f(x_-)}{f(x_+)}.$$

On the other hand, if the derivative of the function $f(t)/g(t)$ is nonnegative at every interior point $a < x < b$ at which it exists, and if every jump of $f(t)/g(t)$ is up, then clearly $f(t)/g(t)$ is weakly increasing on all of $[a,b)$, and so by Lemma 2, $x_H \gtrsim_U x_L$ on $[a,b)$. □

Theorem 7 *(Optimality implies inside domination and uniform outside domination). Let there be N iid students who can work or shirk, and suppose the density assumption holds. Let $\mathcal{P}$ be an optimal absolute grading partition. Then x_L dominates x_H inside $\mathcal{P}$ and x_H uniformly dominates x_L outside $\mathcal{P}$.*

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 3, 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 the inside domination $x_L \gtrsim x_H$ on $[a,b)$.

For the outside domination, 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 3 that

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

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 $\mathcal{P}$. Hence the likelihood ratio property holds for $f_{\mathcal{P}}$ and $g_{\mathcal{P}}$ across consecutive masked intervals. This same logic applies when exam scores are discrete.

The partition $\mathcal{P}$ must consist of intervals, each of which is fine or masked. If $f_{\mathcal{P}}(x)/g_{\mathcal{P}}(x)$ is weakly increasing, then by Lemma 2 we have that $x \gtrsim_U y$ on $\mathcal{P}$. Suppose to the contrary that there is $\alpha < \beta$ with $f_{\mathcal{P}}(\alpha)/g_{\mathcal{P}}(\alpha) > f_{\mathcal{P}}(\beta)/g_{\mathcal{P}}(\beta)$. Then we can assume that either (1) α and β are in the same fine interval, or (2) α is in a fine interval and β is in the next (coarse) interval, or the reverse (3).

Suppose α and β are in the same fine interval. If x_H does not uniformly dominate x_L on $[\alpha,\beta)$, then by Lemma 4 there is a subinterval $[c,d) \subset [\alpha,\beta)$ such that $x_L \succ_U x_H$ on $[c,d)$. But then by Theorem 5, any optimal partition should mask $[c,d)$, a contradiction.

It only remains to consider the case where the drop in $f_{\mathcal{P}}(x)/g_{\mathcal{P}}(x)$ occurs at θ because θ is the cut between a perfectly fine cell $[c,\theta)$ of $\mathcal{P}$ and a masked cell $[\theta,d)$ of $\mathcal{P}$ (or *vice versa*). We argue along the lines of Lemma 4 that $\mathcal{P}$ could not be optimal, because moving the cut from θ to $\theta - \varepsilon$ would lower the payoff to x_L. We rely on the continuity of f to the left of θ, which holds for the generic case or the piecewise differentiable case.

Indeed, the change in expected payoff to L from moving the cut to $\theta - \varepsilon$ is

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

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 $\varepsilon \to 0$, $P(\theta - \varepsilon \leqslant x_H < \theta)$ converges to $\varepsilon f(\theta_-)$, and $P(\theta - \varepsilon \leqslant x_L < \theta)$ converges to $\varepsilon g(\theta_-)$. Thus if $f(\theta_-)/g(\theta_-) > f_{\mathcal{P}}(\theta_+)/g_{\mathcal{P}}(\theta_+)$, 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 $\mathcal{P}$. □

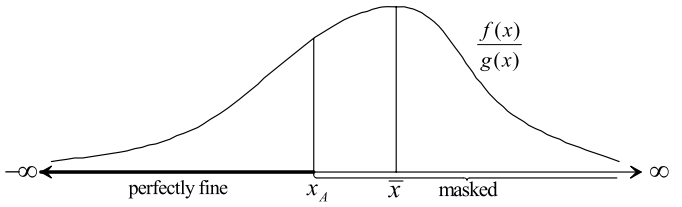

Fig. 5. Normally distributed: $\mu > \tilde{\mu}$, $\sigma > \tilde{\sigma}$.

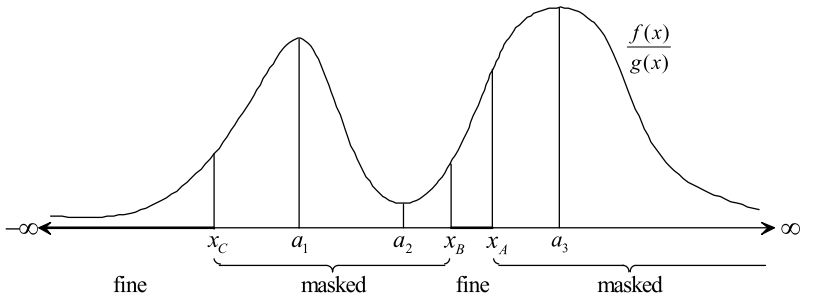

Fig. 6. Generic regular densities.

We can use Theorems 5, 6, and 7 to completely characterize the $\mathcal{A}$-optimal partitions for the normally distributed case, and more generally, for the generic case. Consider again the situation where $f \sim N(\mu, \sigma)$ and $g \sim N(\tilde{\mu}, \tilde{\sigma})$ with $\sigma < \tilde{\sigma}$. We have seen that the function $f(t)/g(t)$ is differentiable and single-peaked, strictly rising for $-\infty < t \leqslant \bar{x}$ and strictly falling for $\bar{x} \leqslant t < \infty$. See Fig. 5. Thus the normal case is differentiable and generic. We know from Theorem 5 that any partition that cuts $(\bar{x}, \infty)$ can be strictly improved by a partition that leaves $(\bar{x}, \infty)$ uncut.

Moreover, since $f(t)/g(t)$ is strictly increasing on $(-\infty, \bar{x})$, we know from Theorem 7 that if there is any cut in $(-\infty, \bar{x})$, say at x_A, then $(-\infty, x_A)$ should be perfectly fine. It follows that the optimal partition must be of the form $\{(-\infty, x_A), [x_A, \infty)\}$ with $(-\infty, x_A)$ perfectly fine and $[x_A, \infty)$ completely masked, and $x_A < \bar{x}$. The point x_A is uniquely defined by the greatest $x \leqslant \bar{x}$ such that

$$\frac{f(x_A)}{g(x_A)} = \frac{P(x_H \geqslant x_A)}{P(x_L \geqslant x_A)}.$$

We argue that the point x_A exists. At $x = \bar{x}$, $f(\bar{x})/g(\bar{x}) > P(x_H \geqslant \bar{x})/P(x_L \geqslant \bar{x})$. As x falls to the left of $\bar{x}$, $f(x)/g(x)$ also falls, but $P(x_H \geqslant x)/P(x_L \geqslant x)$ rises as long as $f(x)/g(x) > P(x_H \geqslant x)/P(x_L \geqslant x)$. Since $f(\bar{x})/g(\bar{x}) > 1$ and $\lim_{x \to -\infty} f(x)/g(x) = 0$, x_A exists.

If $x > x_A$, then the partition $\{(-\infty, x), [x, \infty)\}$ violates the outside condition of Theorem 7, while if $x < x_A$ it violates the inside condition on the cell $[x, \infty)$, as seen by cutting this cell at x_A. The cut at $x = x_A$ preserves both the outside uniform domination, and the inside domination guaranteeing its optimality according to Theorem 6.

The general picture is as follows:

Theorem 8 *(Optimal partitions for generic densities). Consider the case of generic densities f and g. In any optimal partition $\mathcal{P}$, all the cuts are in the rising segments (a_i, a_{i+1}) of f/g, where i is even. If there is a cut in (a_i, a_{i+1}), then the set of all cuts in (a_i, a_{i+1}) is a fine interval $[\alpha_i, \beta_i) \subset (a_i, a_{i+1})$ in P, or else a point $\alpha_i = \beta_i$. In either case,*

$$\frac{f(\beta_i)}{g(\beta_i)} = \frac{\int_{\beta_i}^{y} f(t)\,dt}{\int_{\beta_i}^{y} g(t)\,dt}$$

where y is the smallest cut to the right of β_i, and

$$\frac{f(\alpha_i)}{g(\alpha_i)} = \frac{\int_{x}^{\alpha_i} f(t)\,dt}{\int_{x}^{\alpha_i} g(t)\,dt}$$

where x is the biggest cut to the left of α_i. Thus the optimal partition has at most one more cell than the number of extremal points of f/g.

Proof. Immediate from Theorems 5 and 7, and Lemma 2, using the same logic as in Fig. 5. See Fig. 6 □

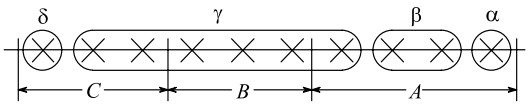

Fig. 7. Absolute vs. relative grading.

5. Grading on a curve with homogeneous students

Though we have only been able to characterize optimal *absolute* grading schemes, at least we can show that they are better than any grading on a curve.

Theorem 9 *(Absolute grading beats grading on a curve). Let there be N iid students who can work or shirk, and suppose the density assumption holds. Let $\mathcal{P}$ be an optimal absolute grading partition. Then $\mathcal{P}$ gives at least as much incentive to work as any grading on a curve, assuming ties are broken randomly.*

The proof relies on the necessity of the inside and uniform outside domination criteria for any $\mathcal{A}$-optimal partition, given in Theorem 7. Starting from an $\mathcal{A}$-optimal partition, we prove the stronger result that conditional on the number of students who get each absolute grade, no grading on a curve will do better.

For the proof we first establish a simple lemma.

Lemma 5. *Denote scores in $[\theta, \infty)$ as A. Suppose $N-1$ students work hard, and each has probability p of getting an A, while one student shirks and has probability $q < p$ 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 less than K/N, while the probability any hard worker got A is more than K/N.*

Proof. The conditional probability the shirker got A is

$$\frac{q\binom{N-1}{K-1}p^{K-1}(1-p)^{N-K}}{q\binom{N-1}{K-1}p^{K-1}(1-p)^{N-K}+(1-q)\binom{N-1}{K}p^{K}(1-p)^{N-K-1}}$$

which is strictly monotonically increasing in q (as can easily be seen by dividing numerator and denominator by the numerator). But when $q = p$, symmetry implies that the expression must be exactly K/N. Hence the probability the shirker got A is less than K/N. Since exactly the proportion K/N students did get A, the probability of the good students getting A must then be more than K/N. □

Proof of Theorem 9. In case the scores are discrete, we can extend each possible score $x \in \{1, 2, \ldots\}$ by the interval $[x, x+1)$, and furthermore assume that an agent who scores x in the discrete model scores uniformly in the interval $[x, x+1)$ in the extended model. The new game thus created is identical to the original discrete game with ties broken randomly. With this extension the discrete case of the density assumption is reduced to the piecewise differentiable case. From now on we shall therefore assume that ties occur with zero probability.

Let $\mathcal{Q}$ be any partition of class rank $\{1, 2, \ldots, |N|\}$, representing an arbitrary grading on a curve.

For any possible distinct exam scores $x = (x_n)_{n\in N}$, and any absolute interval G in $\mathcal{P}$ (whether coarse or fine), let $\mu_G(x)$ be the number of exam scores lying in G. For any curved grade κ, let $\mu_G^\kappa(x)$ be the number of scores in G that also get curved grade κ. Note that since there are no ties, $(\mu_G^\kappa)_{G\in\mathcal{P}}$ can be deduced from $\mu \equiv (\mu_G)_{G\in\mathcal{P}}$. Fig. 7 helps to clarify the situation. In the figure there are 4 A's, 3 B's, and 3 C's on the absolute scale. The curve 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 can deduce that $\mu_A^\gamma = 1$, $\mu_B^\gamma = 3$, and $\mu_C^\gamma = 2$.

Define the join $\mathcal{P} \vee \mathcal{Q}$ of $\mathcal{P}$ and $\mathcal{Q}$ as follows: given scores $(x_n)_{n\in N}$, the exam grade for x_n is strictly higher according to $\mathcal{P} \vee \mathcal{Q}$ than the exam grade for x_m iff *either* the absolute grade for x_n is strictly higher than for x_m, *or* the curved grade for x_n is strictly higher than for x_m. Conditional on $\mu = (4, 3, 3)$, as in the picture, that is achieved by cutting the curved grade cell γ into three curved grades γ_A, γ_B, and γ_C with cardinalities 1, 3, and 2, respectively.

We will now argue that if we grade according to the join $\mathcal{P} \vee \mathcal{Q}$ then the expected exam payoff of the shirker is no more than it was in $\mathcal{Q}$. In fact we will prove more. If any curved grade, such as γ in $\mathcal{Q}$, is refined in $\mathcal{P} \vee \mathcal{Q}$, we show that the expected score of the shirker, conditional on μ and on his being in γ, will not go up. 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 at most zero.

The idea is as follows. Suppose B is a higher absolute grade than C, and both intersect γ. If they were contained in γ, then by the outside uniform domination property of $\mathcal{P}$ (assured by Theorem 7) and by Lemma 5 we would know that the worker is relatively more likely to get B vs. C than is the shirker. Revealing these absolute grades would help the worker and hurt the shirker. If B and C are not contained completely within γ, the same conclusion holds more strongly. Since $C < B$, it is the upper portion of C (and the lower portion of B) that intersects γ. By the inside domination property of $\mathcal{P}$

(assured by Theorem 7) the worker is more likely than the shirker to be in the lower portion of any absolute grade. Thus conditional on being in $\gamma \cap (B \cup C)$, it is still more likely that the worker is in B. Now we present the details.

Let μ be an arbitrary absolute distribution of scores such that $\mu_G \leqslant 1$ for every fine interval G in $\mathcal{P}$. By subdividing fine intervals into smaller and smaller fine intervals, the probability that two scores fall in a single fine interval goes to zero, so we can restrict attention to such μ.

Let γ be any curved grade with $P(x_L \in \gamma|\mu) > 0$. Let $\mathcal{P}^*$ be the collection of intervals G in $\mathcal{P}$ such that $\mu_G^\gamma \geqslant 1$. Clearly $\mathcal{P}^*$ has a finite number of elements.

Let $\tilde{q}$ denote the probabilities of a shirker getting each absolute grade, conditional on the absolute grade distribution μ and the shirker being in γ.

We shall show that if $C < B$ are intervals in $\mathcal{P}^*$ with $\tilde{q}_B > 0$, then $\tilde{q}_C > 0$ and

$$\frac{\tilde{q}_B}{\tilde{q}_C} \leqslant \frac{\mu_B^\gamma}{\mu_C^\gamma}.$$

If $\tilde{q}_B > 0$, then $P(x_L \in B|\mu) > 0$, and from the outside uniform domination property of $\mathcal{P}$ (assured by Theorem 7) and by Lemma 5,

$$\frac{P(x_L \in B|\mu)}{P(x_L \in C|\mu)} \leqslant \frac{\mu_B}{\mu_C} \leqslant \frac{P(x_H \in B|\mu)}{P(x_H \in C|\mu)}.$$

Hence $P(x_L \in C|\mu) > 0$.

If C is fine, then by hypothesis $\mu_C^\gamma = \mu_C = 1$, and obviously $P(x_L \in C \cap \gamma|\mu \ \& \ x_L \in C) = 1$. If the interval C is masked, then by Theorem 7 the shirker dominates inside the cell, hence in either case

$$P(x_L \in C \cap \gamma|\mu \,\&\, x_L \in C) \geqslant \frac{\mu_C^\gamma}{\mu_C}.$$

By Bayes Law used twice,

$$\begin{aligned}\tilde{q}_C &= P(x_L \in C|\mu \,\&\, x_L \in \gamma) = P\big(x_L \in (C \cap \gamma)|\mu\big)/P(x_L \in \gamma|\mu)\\ &= P(x_L \in C|\mu)P(x_L \in C \cap \gamma|\mu \,\&\, x_L \in C)/P(x_L \in \gamma|\mu).\end{aligned}$$

Thus

$$\tilde{q}_C \geqslant \frac{\mu_C^\gamma}{\mu_C}\frac{P(x_L \in C|\mu)}{P(x_L \in \gamma|\mu)}.$$

Similarly, if only the *bottom* part of scores in B are included in γ, then again by using Theorem 7

$$P(x_L \in B \cap \gamma|\mu \,\&\, x_L \in B) \leqslant \frac{\mu_B^\gamma}{\mu_B}.$$

Thus again by using Bayes Law twice,

$$\tilde{q}_B \leqslant \frac{\mu_B^\gamma}{\mu_B}\frac{P(x_L \in B|\mu)}{P(x_L \in \gamma|\mu)}.$$

Therefore

$$\frac{\tilde{q}_B}{\tilde{q}_C} \leqslant \frac{\mu_B^\gamma P(x_L \in B|\mu)}{\mu_B P(x_L \in C|\mu)}\frac{\mu_C}{\mu_C^\gamma} \leqslant \frac{\mu_B^\gamma}{\mu_C^\gamma},$$

as claimed.

It follows that the shirker's expected exam payoff according to $\mathcal{P} \vee \mathcal{Q}$ against the other $\sum_{G \in \mathcal{P}^*} \mu_G^\gamma - 1$ scores in γ must be non-positive. Thus, conditional on μ alone, the expected exam payoff of a shirker is lower when grading by $\mathcal{P} \vee \mathcal{Q}$ than when grading by $\mathcal{Q}$.

Thus we have shown that the expected exam payoff to a shirker is lower under $\mathcal{P} \vee \mathcal{Q}$ than under $\mathcal{Q}$.

To conclude the proof, we need only show that the expected exam payoff of the shirker in $\mathcal{P}$ is even lower (weakly) than his expected exam payoff in $\mathcal{P} \vee \mathcal{Q}$. This follows at once from Theorem 7, which says that conditional on being in a masked cell of $\mathcal{P}$, the score of the shirker dominates the score of a worker. For then, by Lemma 1, any monotonic grading scheme within cells of $\mathcal{P}$ (such as is induced by $\mathcal{P} \vee \mathcal{Q}$) will inevitably work in the wrong direction, weakly increasing the payoff of the shirker. □

6. Work as the unique Nash equilibrium with homogeneous students

Suppose that students' performances are completely independent, regardless of their effort levels (strengthening the conditional independence assumed at the start of Section 4). Fix any monotonic grading scheme. Then the argument given in Section 3.3.1, for the case when i and j are of the same type, carries over mutatis mutandis to the setting of homogeneous students, since it relies only on the symmetry between i and j. Thus, again, whenever all work is a strict NE, it is the unique NE.

7. Games of separable status with incomplete information

So far we have considered games of complete information. Every student in the class knows the characteristics of all the others. One might wonder if our analysis can be extended to games of incomplete information, in which each student knows his own characteristic, but has a probability distribution on those of others. If there is a common prior, and if the exam score of each student is independent of the effort levels and scores of the other students (as we have often assumed), and if there is absolute grading, then the answer is yes, because the exam payoff of each student is "separable" across his rivals, in a sense to be made precise shortly. We shall show that, in this scenario, any finite-player game of incomplete information is strategically equivalent to a continuum-player game of complete information. This equivalence is useful because the continuum games are easy to analyze.

Suppose as before that there is a set $N = \{1, \ldots, N\}$ of players, each of whom can be one of a finite number of types $t \in T$. Previously we assumed that the type of each player was commonly known. Now we suppose that they are all drawn i.i.d. from T with probability (common prior) μ (so $\sum_{t\in T} \mu(t) = 1$).

Let γ be an absolute grading scheme. Let E_t be the action (effort) space of agents of type $t \in T$, as before. Given two players of types $t \in T$ and $s \in T$, who are choosing actions $e_t \in E_t$ and $e_s \in E_s$, define

$$u_\gamma^{t,s} : E_t \times E_s \to u_\gamma^{t,s}(e_t, e_s) = \text{Prob}(x_t >_\gamma x_s | t, e_t, s, e_s) - \text{Prob}(x_s >_\gamma x_t | t, e_t, s, e_s)$$

where x_t, x_s are the stochastic scores resulting from effort levels e_t, e_s, and, recall, "$x_t >_\gamma x_s$" means that γ awards a higher grade to x_t than to x_s etc. In games of (additive) status the payoff of a player is the aggregate of his payoffs against all his rivals. More precisely, let players $1, \ldots, N$ of types $t = (t_1, \ldots, t_N)$ choose actions $e = (e_1, \ldots, e_N)$. The *ex post* payoff of player n is

$$U_\gamma^n(e, t) = u_\gamma^n(e, t) - e_n = \left[\sum_{j\in N\setminus n} u_\gamma^{t_n,t_j}(e_n, e_j) \right] - e_n.$$

This defines a game of incomplete information $G_\gamma = (N, T, (E_t)_{t\in T}, \mu, (U^n)_{n\in N})$. Our equivalence theorem below depends on the *separability* property:

$$u_\gamma^n(e, t) = \sum_{j\in N\setminus n} u_\gamma^{t_n,t_j}(e_n, e_j)$$

and not on any of the special features of $u_\gamma^{t_n,t_j}(e_n, e_j)$.

We define a Bayesian Nash equilibrium (BNE) to be an (*ex ante*) symmetric Nash equilibrium (NE) of the game just described.

Using the fact that players are iid and are all choosing the same strategy, we see that $\tilde{e} : T \to \bigcup_{t\in T} E_t$ is a BNE if, $\forall t \in T$

$$\tilde{e}(t) \in \arg\max_{a\in E_t} \left\{ (N-1) \left[\sum_{s\in T} \mu(s) u^{t,s}\big(a, \tilde{e}(s)\big) \right] - a \right\}.$$

Now consider instead the finite-type continuum-player game of complete information defined as follows. There are disjoint intervals $I_t, t \in T$, of Lebesgue (population) measure $\lambda(I_t) = (N-1)\mu(t)$, representing players of type t, with action set E_t. Given any measurable function $\tilde{e} : \bigcup I_t \to \bigcup E_t$ with $\tilde{e}(x) \in E_t$ whenever $x \in I_t$, define the payoff to any agent $x \in I_t$ by

$$\sum_{s\in T} \int_{I_s} u^{t,s}\big(\tilde{e}(t), \tilde{e}(s)\big)\, d\lambda(s).$$

Theorem 10. *BNE of the finite-player incomplete information game are in one-to-one correspondence with the type-symmetric NE of the continuum-player complete information game.*

Proof. The one-to-one correspondence is evident from the argmax display above, interpreting $\tilde{e}(t)$ in the BNE as a constant function on I_t in the NE of the continuum game. □

Table 1
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	0695262507	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

We now apply Theorem 10 to the case of disparate students described in Section 3. (The case of homogeneous students does not allow for incomplete information, since an agent who knows his own type and knows that all the students have to be of the same type as him must know the type of every student.)

Suppose there are N students, each of whom can be of disparate type $1, \ldots, \ell$, with probability $\mu(1), \ldots, \mu(\ell)$. Each student is informed of his type, but not of the others'. The number of students N_i that turn out to be type i is now random, and unknown to them and to the professor. What is the best absolute grading scheme? Theorem 10 applies, since the spearability hypothesis holds under absolute grading. By Theorem 10, we need only analyze the continuum player game with complete information. The analysis of this game is given by exactly the same formulas displayed in Section 3.3 for the finite player game of complete information except that *both* N_i and $N_i - 1$ need to be replaced by $\mu(i)$. The analysis of Section 3.3 applies simply and directly for the continuum player game, without any need to approximate the continuum with finite-player sets.

Equilibrium of the incomplete information game is thus easy to compute. Consider $\ell = 20$, and fix the measures $\mu_1 = \mu_2 = \cdots = \mu_\ell = 1$ of being of any type $1, \ldots, \ell$. For $\ell = 20$, the incentive $I_\ell \approx I^*$ is about 1.389 for each student. Since $p_1 = 1$, I^* is also the measure of students receiving the lowest grade $i = 1$: $I^1(\bar{p}) = \bar{p}_1(2 - \bar{p}_2) = (2 - \bar{p}_2) = 1 + (1 - \bar{p}_2)$.

In Table 1 we list the optimal $(p_1, \ldots, p_{20})$ and the measure of students for each grade $i = (1, \ldots, 20)$.

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 bottom grade (aside from the failing grade $i = 0$ that nobody gets) $i = 1$ is the most commonly given.

8. Concluding comments

8.1. Heterogeneous students

Consider the situation in which students are neither identical nor disparate. To be concrete, suppose that an exam consists of K questions. Each student $n = 1, \ldots, N$ 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 5. 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 $N = 10$ in our working paper Dubey and Geanakoplos (2005). We show by simulation that the coarse partition $\{[0, 9), [9, 18), [18, 27)\}$ into just three grades gives greater minimum incentive than the perfectly fine partition $\{[0], [1], [2], \ldots, [27]\}$.

8.2. Midterms

The introduction of midterms before a final often makes it even more advantageous to have coarse grading. Even the coarsening achieved by first averaging the numerical scores of the midterm and the final, and then clumping these averages into letter grades, may not suffice. It might become necessary to clump midterm scores into grades, and to clump final scores into grades, and then to average the grades.

This point is illustrated in a detailed example in our working paper Dubey and Geanakoplos (2005).

References

Cole, Harold L., Mailath, George J., Postlewaite, Andrew, 1992. Social norms, savings behavior, and growth. J. Polit. Economy 100 (6), 1092–1125.
Cole, Harold L., Mailath, George J., Postlewaite, Andrew, 1995. Incorporating concern for relative wealth into economic models. Fed. Reserve Bank Minneapolis Quart. Rev. 19 (3), 12–21.
Cole, Harold L., Mailath, George J., Postlewaite, Andrew, 1998. Class systems and the enforcement of social norms. J. Public Econ. 70, 5–35.
Corneo, Giacomo, Jeanne, Oliver, 1997. On relative wealth effects and the optimality of growth. Econ. Letters 54, 87–92.
Demougin, Dominique, Fluet, Claude, Helm, Carsten, 2006. Output and wages with inequality averse agents. Can. J. Econ. 39, 399–413.
Direr, Alexis, 2001. Interdependent preferences and aggregate saving. Ann. Econ. Statist. 63–64, 297–308.
Dubey, P., Geanakoplos, J., 2004. Grading exams: 100, 99, ..., 1 or A, B, C? Incentives in games of status. Cowles Foundation discussion paper No. 1467, Yale University.
Dubey, P., Geanakoplos, J., 2005. Grading in games of status: Marking exams and setting wages. Cowles Foundation working paper #1544.
Dubey, P., Geanakoplos, J., forthcoming. How to pay workers when wages also confer status. Cowles Foundation discussion paper.
Duesenberry, James S., 1949. Income, Saving and the Theory of Consumer Behavior. Harvard University Press, Cambridge.
Easterlin, Richard, 1974. Does economic growth improve the human lot? In: David, Paul A., Reder, Melvin W. (Eds.), Nations and Households in Economic Growth: Essays in Honor of Moses Abramowitz. Academic Press, New York, pp. 87–125.
Englmaier, Florian, Wambach, Achim, 2006. Optimal incentive contracts under inequity aversion. Unpublished job market paper, Harvard Business School.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition and cooperation. Quart. J. Econ. 114, 817–868.
Frank, Robert H., 1985. Choosing the Right Pond: Human Behavior and the Quest for Status. Oxford University Press, New York.
Hopkins, Ed, Kornienko, Tatiana, 2003. Ratio orderings and comparative statics. Working paper, University of Edinburgh.
Itoh, Hideshi, 2004. Moral hazard and other-regarding preferences. Japanese Econ. Rev. 55 (1), 18–45.
Moldovanu, B., Sela, A., Shi, X., 2007. Contests for status. J. Polit. Economy 115, 338–363.
Pollak, Robert, 1976. Interdependent preferences. Amer. Econ. Rev. 66 (3), 309–320.
Robson, Arthur, 1992. Status, the distribution of wealth, private and social attitudes to risk. Econometrica 60 (4), 837–857.
Shaked, M., Shanthikumar, J.G., 1994. Stochastic Orders and Their Applications. Academic Press, San Diego.
Spence, M., 1974. Market Signaling: Informational Transfer in Hiring and Related Processes. Harvard University Press, Cambridge.
Veblen, Thorstein, 1899. The Theory of the Leisure Class. Macmillan, New York.