Empir Econ (2013) 44:1563–1592
DOI 10.1007/s00181-012-0578-0

Grades as incentives

Darren Grant · William B. Green

Received: 16 January 2011 / Accepted: 12 January 2012 / Published online: 1 April 2012

Abstract This paper examines how grade incentives affect student learning across a variety of courses at two universities, using for identification the discrete rewards offered by the standard A–F letter-grade system. We develop and test five predictions about the provision of study effort and the distribution of numerical course averages in the presence of the thresholds that separate these discrete rewards. Surprisingly, all are rejected in our data. There is no evidence that exam performance is improved for those students that stand to gain the most from additional study.

Keywords Educational assessment · Thresholds · Behavioral incentives

JEL Classification I21 · A22 · D10

1 Introduction

Effective learning requires initiative by students as well as teachers. Yet, while performance incentives for teachers have received considerable attention from economists (and the general public), less attention has been paid to motivating students, and most recent work in this vein has focused on financial incentives (Angrist et al. 2009; Jackson 2010; Fryer 2010), which are rare in practice. Few studies have examined the most fundamental incentive offered to students, grades, despite their prevalence, importance, and interesting economic properties. With a thorough theoretical and empirical analysis of the incentive properties of grades, this paper helps fill that gap.

D. Grant (✉) · W. B. Green
Department of Economics and International Business, Sam Houston State University, Huntsville, TX, USA
e-mail: dgrant@shsu.edu

W. B. Green
e-mail: eco_wbg@shsu.edu

A brief sampling of the literature illustrates the multifaceted value of such an inquiry. Farkas and Hotchkiss (1989) consider grade incentives to be one component of comprehensive educational reform, in which states raise educational standards and schools grade more rigorously to encourage students to meet those standards. Bonesrønning (2004), in contrast, treats the strength of grade incentives as an identifiable, mutable determinant of teacher effectiveness. Oettinger (2002) focuses on the student response to these non-monetary incentives, while Babcock (2010) adopts a broad historical perspective, asking whether collegiate grade inflation helps explain the downward trend in study time. At the level of the student, teacher, school, and system, it is useful to know how grade incentives affect student learning.

Our empirical methodology relies on a curious feature of the typical American grading system: the presence of thresholds, which divide grades into discrete units of A, B, C, D, and F. (The plus/minus grading system adopted by some universities does not apply in our data.) In the neighborhood of these thresholds the expected marginal benefit of increased study effort is highly non-monotonic. In consequence, several detailed predictions about effort provision and the distribution of final course averages can be generated. Because these predictions are across students within the same class taught by the same instructor, they obviate the need to control for the instructor's teaching methods, whose absence in cross-teacher analyses is a possible source of bias.

Using simple non-parametric methods, we test these predictions and infer the effect of grade incentives on learning at all four grade thresholds across the full distribution of student motivation for several college courses at two universities. The results are consistent throughout: stronger incentives are not associated with improved performance. Because studies cited below indicate that study effort is associated with improved exam performance, this suggests that grades are not effective motivators on the margin, at least for tests in college classes at universities comparable to those studied here.

2 Grades and incentives: three perspectives

The natural statistical approach for analyzing grades' incentive effects is to assume the null that there are no such effects, and determine whether it can be rejected by the data. While this null runs counter to standard economic theory, it should not be considered a foil, destined to fail. Its plausibility is suggested by two alternate perspectives on the effects of grade incentives: the historical, which inquires whether the letter-grade system was adopted in order to motivate students, and the psychological, which identifies those motivators that appear to have the strongest effects on learning.

2.1 Historical perspective

Educational assessment originated toward the end of the medieval period, in order to group students within schools on the basis of mastery. It spread across Europe throughout the Renaissance, as the state tried to improve the quality of its civil servants, who were increasingly selected on the basis of merit instead of social class (Wilbrink 1997).

In America, the first defined scale for differentiating students appeared at Yale in 1785, using four tiers, as in English universities. In 1813, Yale moved to a four-point numerical scale that included both whole numbers and decimals, while Harvard contemporaneously adopted a twenty-point scale. Throughout the remainder of the nineteenth century American universities experimented with a variety of marking systems, including written reports, adjectives such as "good" and "exemplary," and a variety of numerical scales, often quite detailed (Smallwood 1935).

Modernization of the curriculum toward the end of the nineteenth century seems to have brought with it the first letter grades: a five-tiered, A through E system instituted at Harvard in the 1880s. This system was explicitly intended to *diminish* motivation:

> The Faculty last year did away with the minute percentage system of marking, and substituted a classification of the students in each course of study in five groups, the lowest of which includes those who have failed in the course. It is hoped that this grouping system will afford sufficient criteria for the judicious award of scholarships, honorable mention, and the grade of the Bachelor's degree, while it diminishes the competition for marks and the importance attached by students to College rank in comparison with the remoter objects of faithful work (*Annual Report of the President of Harvard* 1885, p. 9, quoted in Smallwood 1935, p. 51).

As Harvard's new curriculum and teaching methods spread throughout American higher education, so did its new grading system.[1]

A similar shift occurred in American public schools during this period. As enrollment and professionalism increased dramatically, assessment initially evolved away from written narratives toward percentages on examinations in different subject areas. Then Wisconsin researchers Daniel Starch and Edward Charles Eliot (1912, 1913) challenged the reliability of percentages as indicators of achievement, showing that teachers assigned a wide variety of grades to identical papers, with percentage scores ranging at least thirty-four points in English and as much as sixty-seven points in math. In response, schools moved away from percentage scores to fewer, larger categories, such as the "Excellent," "Good," "Average," "Poor," and "Failing" system that presaged today's A–F scale.

In summary, grading systems evolved with the educational system, partly in response to demands for better information about student performance, but were not explicitly designed to motivate students. This holds in particular for the introduction of thresholds, used first to group students into homogenous classes and expedite cost-effective instruction, then adopted by Harvard to weaken "competition for marks," and finally employed by the followers of Starch and Eliot to mask the disparity in instructors' grading standards.

[1] Grant (2010) shows that, compared to a continuous incentive such as a numerical grading system, a threshold can increase or decrease average effort, depending on its placement. Even when effort is diminished on average, however, it is increased for those students on the border between two letter grades. The marginal rewards for improved performance are reduced for infra-marginal students and enhanced for marginal, or borderline, students.

2.2 Psychological perspective

Research in educational psychology, summarized in Stipek (1996) and Elliot and Zahn (2008) and applied to economics in Hadsell (2010), also casts doubt on grades' motivational efficacy.

This literature initially emphasized the Skinnerian model, in which behavior responded to "extrinsic" reinforcements, such as trinkets, money, or grades, and in which these reinforcements could be adjusted in an almost Keynesian way to bring about desired outcomes. But over time this model has been de-emphasized in favor of a broader model that also allows internal, or "intrinsic," motivations, and which mediates the effect of external reinforcements through cognitions that influence the way in which students respond to incentives and their objectives in doing so.

This research concludes that extrinsic motivation and intrinsic motivation are substitutes: students have an intrinsic "achievement motive" that is *weakened* by the use of incentives. This diminishes the potency of extrinsic rewards. Furthermore, extrinsic incentives' effects are influenced by students' perceptions of competence and self-efficacy. If these are poor, students adopt a "performance-avoidance" goal—essentially a maximin objective that tries to moderate bad outcomes rather than strive for good ones. When this happens, incentives' effects are yet further diminished.

These findings are buttressed by a broader literature in psychology and, more recently, behavioral economics (Amabile 1983; Vendantam 2008; Ariely et al. 2009; Fryer 2010; Grove and Hadsell 2011, in education) on the value of extrinsic incentives. The general finding is that they improve performance in "algorithmic," or repetitious, tasks but are less effective or even counterproductive at "heuristic" tasks that require creativity, concentration, or intuition. Because learning has generally been classified as heuristic, this too suggests that extrinsic grade incentives may not be effective motivators.

2.3 Economic perspective

As it stands, though, few studies directly explore the effects of grade incentives on student achievement. In the education literature grades are more commonly analyzed in the context of assessment, not incentives (such as in the journal *Studies in Educational Evaluation*); when grade incentives are discussed, it is usually as an inducement for students to participate in the educational intervention that is the focus of the study. These incentives often achieve their purpose (for example, Grove and Wasserman 2006), but if participation in these interventions is algorithmic, while learning is essentially heuristic, then this success need not mean that grades motivate students to learn on their own. This research leaves that question, the focus of this paper, largely unaddressed.

Thus, most studies of grades' incentive effects have been done by economists. Most relate cross-instructor variation in grading standards to study time or learning outcomes (see the review by Bishop 2006; Betts and Grogger 2003; Bonesrønning 1999, 2004; Figlio and Lucas 2004; Grant 2007; Iacus and Porro 2008; Merva 2003). More difficult instructors do have better learning outcomes, but this might have more to do with teaching methods, which are not controlled for, than incentives. While

no study directly relates teachers' grading standards to their pedagogical skill, it is generally accepted that high expectations for achievement are an important component of teacher effectiveness. If such expectations are associated with high grading standards and with pedagogical skill, the effect of such standards on achievement is biased upward in studies that do not account for the effectiveness of the teacher's instructional methods. The same concern applies, to a smaller degree, to Babcock (2010), which relies on within-teacher, cross-class variation, and finds that higher grading standards increase student study time.

Because pedagogical skill is difficult to measure, the only practical way to fully address this concern is to identify incentives' effects using within-instructor, within-class variation. This is achieved by Oettinger (2002), who explores how grade thresholds affect final exam performance for college students, as we do. He concludes that they matter. Oettinger's study design is similar to ours, though our theory places more emphasis on uncertainty and our analytical approach is somewhat less structured; the types of students studied may also differ. Below we compare Oettinger's estimates to ours, and argue there is less dissonance between them than appearances suggest.

In summary, most evidence supporting the effectiveness of grade incentives is potentially subject to omitted variable bias. Furthermore, the historical record and the psychology literature suggest these incentives may, in fact, be ineffectual. This null hypothesis is a legitimate possibility.

3 The behavioral effects of grade thresholds

Many courses above the elementary grades assess performance numerically, by taking an average of scores on tests, homework, quizzes, etc., and then use a grading scale to convert this "final average" into letter grades. This process has two noteworthy features. First, measurement is imperfect: most tests only sample the information covered in the course, so that the score on that test measures each student's "true" knowledge with error.[2] Second, the grading scale converts a continuous measurement, the final average, into a set of discrete rewards, letter grades of A, B, C, D, or F. If measurement were perfect, the marginal benefit of improved performance would be nil until one is about to cross the threshold, after which it is nil again. Since it is imperfect, expected marginal benefits are continuous but non-monotonic, rising, then falling in the neighborhood of the threshold.

These features facilitate hypothesis testing in this paper, which examines how grade incentives affect final exam performance. Because individuals vary in their pre-exam proximity to the numerical threshold separating one letter grade from another, the expected returns to additional study effort also vary. As a result, we can generate several predictions about how final exam performance varies with proximity to this threshold, and concomitant predictions about how the density of final course averages

[2] While this "sampling error" suffices to motivate the model used here, it is not the only source of uncertainty in the returns to study effort captured by the model: the student may not perfectly anticipate the effectiveness of study time, the content or difficulty level of the final exam, etc. As long as these expectations are reasonable, the model below, and predictions derived therein, will continue to hold. On the other hand, there is evidence that collegiate business students overestimate their likely exam performance (Grimes 2002; Clayson 2005).

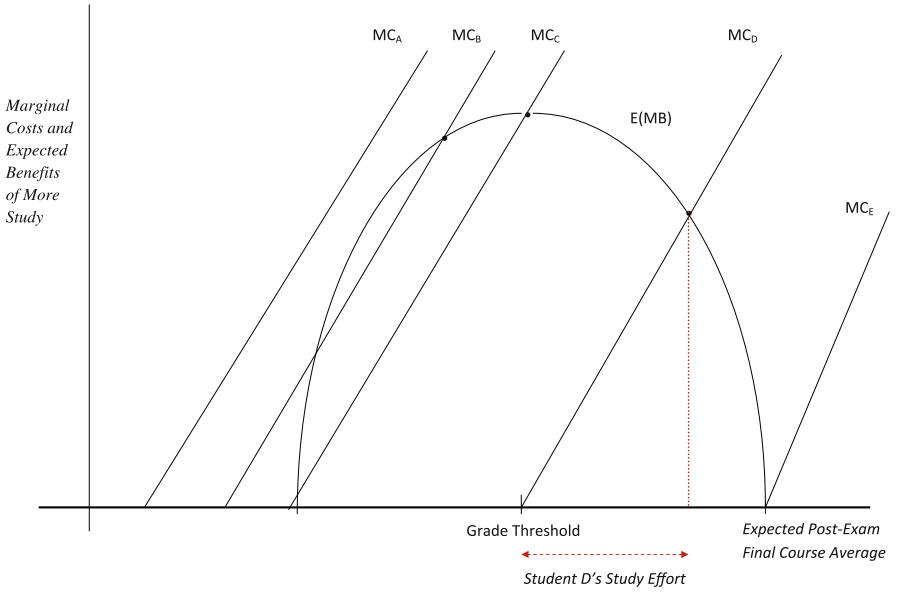


Fig. 1 Analysis of the amount of study effort exerted, conditional on the pre-exam average

varies with proximity to the threshold. These predictions and their normative implications are formally developed for a general model of thresholds in the companion paper, Grant (2010), which applies to a wide range of empirical phenomena, as discussed therein. Here they are laid out with a simple graphical model, depicted in Figs. 1 and 2, and justified intuitively.[3]

This model permits some baseline level of study effort during the semester, but assumes that study effort for the final exam is "strategically" determined, as a function of its marginal costs and its expected marginal benefits. The latter, the parabola in Fig. 1, are centered around the threshold, reflecting the error in measuring performance discussed above, which is assumed to be distributed symmetrically around zero. Marginal costs, the upward sloping lines in the figure, are represented for five students with different pre-exam averages, ranging from student A, a weak student whose pre-exam average is far below the threshold, to student E, a strong student whose pre-exam average is far above the threshold.[4] Absent strategic study effort, each student's expected final course average is indicated on the horizontal axis. Expending effort increases the student's expected exam score, thus increasing their chances of

[3] This model, which focuses on uncovering testable implications of threshold incentive effects, is neither the most general model of study effort available (which would probably be Bishop 2006) nor appropriate when grades are assigned on a curve (see Becker and Rosen 1992). Elements of the model presented here are found in each of these papers, and in Oettinger (2002).

[4] These marginal costs will include the value of time spent in alternative uses, possibly including studying for other final exams. The more exams, the higher and flatter marginal costs will be. This will diminish the time a student spends studying for exams in all classes, borderline and not, and the marginal returns to study effort across classes will tend to be equalized. The difference in study effort between borderline and non-borderline students within a class may also diminish, but is not guaranteed to do so.

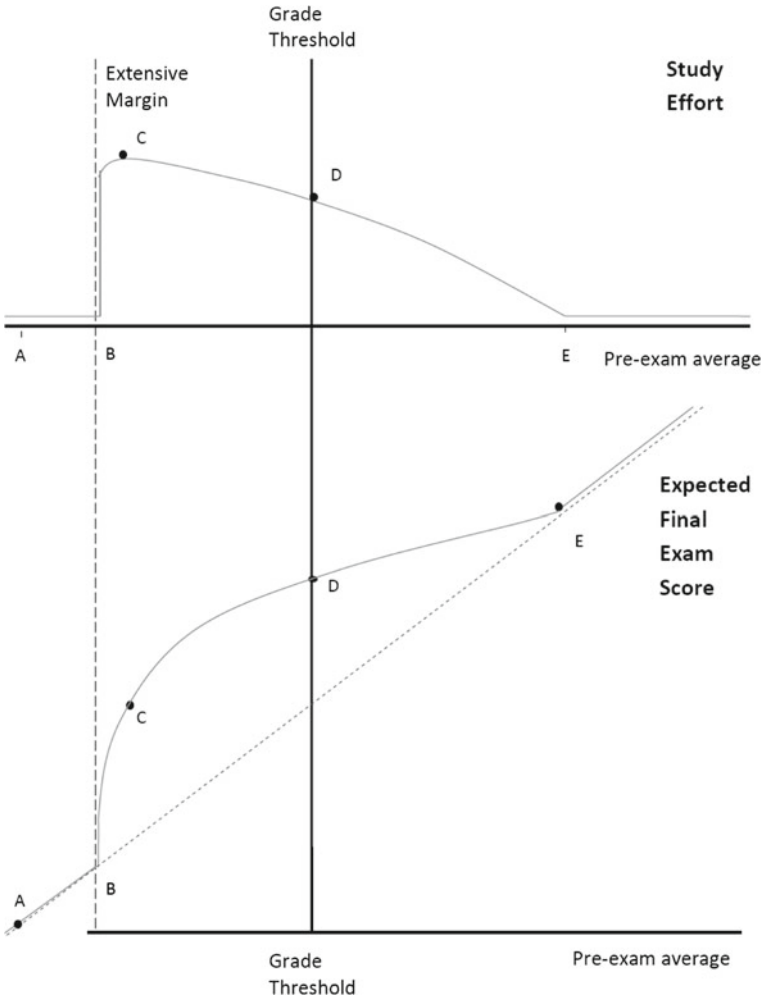


Fig. 2 *Top* pre-exam average—effort locus. *Bottom* pre-exam average—expected final exam score locus

getting the higher grade. When the vertical axis is measured in logs, this graphical model closely resembles the formal model in the companion paper.

Student A is too far below the threshold to make additional study worthwhile: marginal costs never intersect expected marginal benefits. Even when they do intersect there may be only a local maximum, as for student B, who resides just below the extensive margin at which the *total* benefits of effort equal the *total* costs. For this student, the accumulated surplus of marginal costs over marginal benefits, left of the lower point of their intersection, just exceeds the accumulated surplus over marginal benefits over marginal costs, between the lower and upper points of intersection. Thus, optimal effort is still zero. This is no longer true to the right of the extensive margin, between students B and E. Effort becomes positive, determined at the intersection of

marginal costs and expected marginal benefits. Moving right from student B, effort can initially rise and will eventually fall with the pre-exam average. Eventually, for student E and beyond, it returns to zero.

Mapping out the study effort implied by this model as a function of the pre-exam average yields the graph at the top of Fig. 2; the implied relation between the pre-exam average and the expected final exam score is given at the bottom of Fig. 2. These graphs illustrate three predictions of the model, each of which can be described heuristically.

1. **Peak effort property:***Students whose pre-exam averages are far below the threshold dividing two letter grades put forth little effort; those near it put forth more; those in between put forth the most.* This property, which establishes the non-monotonicity of effort, has been noted by several other researchers.
2. **Sawtooth property:***Effort rises more quickly than it falls; that is, in absolute value, the slope of line BC in Fig. 2 (top) is greater than the slope of line CE. Thus, the relation between the pre-exam average and final exam effort takes a sawtooth shape.*This follows both from the existence of the extensive margin, at which effort increases discretely, and from the geometry of Fig. 1. (The point of intersection responds more to increases in the pre-exam average when marginal costs and expected marginal benefits are more similarly sloped, which occurs to the left of point C.)
3. **Stair step property:***More able students have better expected outcomes than less able students.* The relation between the pre-exam average and the final average always slopes upward, like the rounded stair step at the bottom of Fig. 2. Intuitively, as long as there is increasing marginal disutility of study effort, a higher pre-exam average "endowment" will be "spent" partly on reduced study and partly on increased chances of receiving the higher grade.

Study effort depends on the student's proximity to the nearest grade threshold, in a predictable, non-linear fashion. Across the full spectrum of course grades, this pattern should be repeated at each threshold.

All of these predictions can be tested qualitatively, by re-creating Fig. 2 empirically from data on students' pre-exam averages, final exam scores, and post-exam final averages—that is, by forming non-parametric estimates of the relationships between these variables—and seeing if they match the patterns described above. These estimates will be precise if the data include many students, either from one very large class or from several smaller sections of the same subject taught by the same instructor. We adopt the latter approach, amalgamating ten or more such sections over a multi-year period and estimating these relationships with a non-parametric regression, specified below, that includes dummy variables to distinguish the different semesters in which the instructor teaches the class. For some results, we go even further and pool all instructors together, yielding more than 3,000 observations.

With this non-parametric regression we can also formally test for threshold incentive effects. [5] These effects will be absent if students are only motivated intrinsically;

[5] The methodology that ensues is an example of "measuring the affinity of parametric and nonparametric models," of which much as been written—see Pagan and Ullah (1999, Section 3.13), Yatchew (1998), or the extensive application to thresholds in the companion paper.

then they should approach the final exam as they have approached all the other tests. In that case the relation between students' final exam scores, F, and their pre-exam averages, P, should be adequately represented by a simple parametric "trend." In practice our trend is quadratic,[6] yielding the following specification:

$$F_{s,t} = \beta P_s + \gamma P_s^2 + \delta_t + \varepsilon_{s,t} \tag{1}$$

where s indexes students and t time (semester*year), and δ represents the aforementioned semester*year dummy variables (whose presence makes a constant term superfluous).

This is the null hypothesis. The alternative is that this simple parametric relation is inadequate, because strategic effort generates systematic deviations of F from trend:

$$F_{s,t} = g\,(P_s) + \delta_t + \varepsilon_{s,t} \tag{2}$$

with $g(P)$ a nonparametric function that conforms to the properties listed above. The null hypothesis implies $g(P) = \beta P + \gamma P^2$. Using a loess smoother in SAS procedure GAM, we estimate this partially linear model and test the significance of this null.[7] The estimates of expected performance, $\hat{g}$(P), and of the effort incentivized by the threshold, $\hat{g}$(P) – $\hat{\beta}$P- $\hat{\gamma}$ P^2, suffice to form graphs like Fig. 2.

Thus this empirical approach, in addition to being simple and nonrestrictive, is also descriptive. The effect of any study effort motivated by grade incentives, measured as the average percentage point improvement on the final exam, can be read right off the graph. In the companion paper this method clearly identifies strong incentive effects, adhering to the properties above, from a somewhat unusual threshold: the finish time in an ultramarathon required to win a coveted medal.

Incentive effects can also be revealed in the distribution of final course averages. Basic intuition and Fig. 2 generate a fourth testable prediction.

4. **Bunching:** *The empirical density of final averages in a modest interval just above the grade threshold should exceed that in an interval of equal size just below the threshold.*

This can be tested using "the caliper method" (explicated in Gerber and Malhotra (2008); implemented in economics by Borghesi (2008), and others; and extended below to transition rates), simply by calculating these empirical densities and testing the null that they are equal.

A corollary of this property yields a final prediction.

[6] The pre-exam average is also affected by random measurement error. That is, the earlier tests in the course are also imperfect indicators of students' "true" performance, which would be the ideal independent variable. Thus, the pre-exam average suffers from errors-in-variables bias, which makes its coefficient smaller than one. This bias need not be constant, however, because this error is smaller at high grades and larger at low grades (as shown in the scatterplots in the Appendix). Thus, a slight convex shape is expected, and appears for three of the five instructors in the data.

[7] For the figures, the bandwidth is chosen to cleanly resolve the smallest feasible scale over which threshold incentives might be expected to be observed, 2 percentage points in width. This bandwidth under smooths somewhat compared to the bandwidth chosen using generalized cross validation, which is adopted for the formal hypothesis tests in Table 1.

5. **Asymmetric transitions:** *Transitions between the pre-exam and post-exam course averages should be asymmetric, with more individuals near the threshold moving slightly up instead of slightly down.*

Again these transitions can be measured empirically, and the null that the two transition rates are equal easily tested.

The regression tests and the distributional tests are complementary, because they embody different student expectations about the difficulty of the final exam. The former implicitly assume that students expect the final exam to be no more or less difficult than the tests that preceded it: perfect (though perhaps reasonable) myopia. The latter implicitly assume that students know exactly how hard the final exam will be: perfect foresight. Here, both methods yield the same conclusion.

4 Data

The data used in this analysis were generously provided by five university instructors teaching four different courses, both upper and lower division, at two Texas universities during various subsets of the years 1998–2007. The courses, Principles of Accounting, Principles of Microeconomics, Business Statistics, and a "Business Analysis" course combining elementary calculus and probability concepts, are all required for a bachelor's degree in business at their respective universities. Summary details about the courses, instructors, and grading policies are found in Table 1.

Typically, university grading systems are either norm-referenced or criterion-referenced. In the former students are evaluated relative to one another; thresholds still separate letter grades, but are not specified in advance, and so cannot motivate students much on the margin. In contrast, criterion-referenced grading sets absolute standards, using the philosophy that grades should reflect mastery of specific course material. In these systems, thresholds are expected to incentivize effort as previously outlined. All instructors in our sample use criterion-referenced grading on the standard scale, in which 90 % is an A, 80 % a B, 70 % a C, 60 % a D, and below 60 % an F. No plus/minus grades are possible.

We drop all students whose course average, prior to taking the final exam, is less than 50 %. These students are almost destined to fail the course. For all remaining students, our data record all test scores and homework grades, along with the formula used to compute each final course average, which is also given to students in advance on the course syllabus. We can thus compute the student's pre- and post-exam course averages, as can the student herself. All instructors evaluated students, primarily or wholly, on the basis of two to four midterm exams and a final examination that was, except for one instructor, mandatory. (Homework counted for 10 % of Instructor 5's final average, less for the other instructors.) Generally the final exam was worth about one-quarter of the final average. Most exams, including the final, consisted of multiple choice questions, occasionally supplemented with short answer questions or problems requiring simple calculations or graphing. No instructor tried to make the final more difficult than the other tests. Three of the five, in fact, constructed their final exams by modifying slightly, sometimes very slightly, questions asked on previous tests.

Table 1 Course characteristics and sample sizes

	Instructor 1	2	3	4	5
Course taught	Business analysis	Principles of micro	Principles of micro	Business statistics	Principles of accounting
University where taught	SHSU	UTA	SHSU	SHSU	SHSU
Grade level of course	Sophomore	Sophomore	Sophomore	Junior	Sophomore
Sample size	1132	655	943	704	856 total 468 take final
Marginal students (see note)	203	205	226	184	—
Grading scale	90, 80, 70, 60	90, 80, 70, 60	90, 80, 70, 60	90,80, 70, 60	90, 80, 70, 60
Grading system	Criterion-referenced	Criterion-referenced	Criterion-referenced	Criterion-referenced	Criterion-referenced
Adjust points on borderline?	A little	A little	A little	A little	About three points
Contribution of final exam to final grade	20–25 %	25–40 %	25 %	15–50%	0–20 %
Years full-time teaching exp. in 2007	13	12	36	9	16
Final exam mandatory?	Yes	Yes	Yes	Yes	No
Test/exam format	MC, problems	Problems, MC, short answer	MC, short answer	MC, problems	MC, problems
Sample period	2002–2007	2004–2007	1998–2007	2002–2007	2005–2007

Instructor 3 allows students to drop any test except the final exam, and Instructor 4 allows the final exam to replace the lowest test, which it almost always does. In the empirical work, the pre-exam average for these instructors' students accounts for this dropped test. Marginal students' final grades would change with a five-point change, up or down, on the final exam. *MC* Multiple choice

For each instructor, exam difficulty varies modestly across semesters, with the semester dummies in our regressions having a standard deviation of about 3 percentage points.

There is nothing atypical about these course characteristics; nor is there anything atypical about the universities at which these courses were taught: Sam Houston State University, a public, seventeen-thousand student, *U.S. News* third-tier regional university; and the University of Texas at Arlington, a public, twenty-five thousand student, U.S. News fourth-tier national university. Median incoming SAT scores at both schools modestly exceed the national average of about 1,020; 6-year graduation rates, around 40 %, are typical for universities of this type. We do not claim that students in all universities behave as these students do, only that these universities are not unrepresentative of the higher education system in the United States.

The instructors in our data are all terminally qualified, currently possessing a century of combined full-time teaching experience; in their first year in our sample each has at least 4 years prior experience teaching that course. Course evaluations and administrators' judgements suggest that these instructors typically are successful in teaching these courses and that they set appropriate course expectations and grading standards. Each instructor in our data teaches more than 650 students, so that both parametric and nonparametric estimates of effort provision, as reflected in final exam scores, can be obtained with reasonable precision. In fact, the standard errors on the empirical results presented below are about 1 percentage point, so that improvements in final exam performance of only 2 or 3 percentage points can be distinguished statistically.

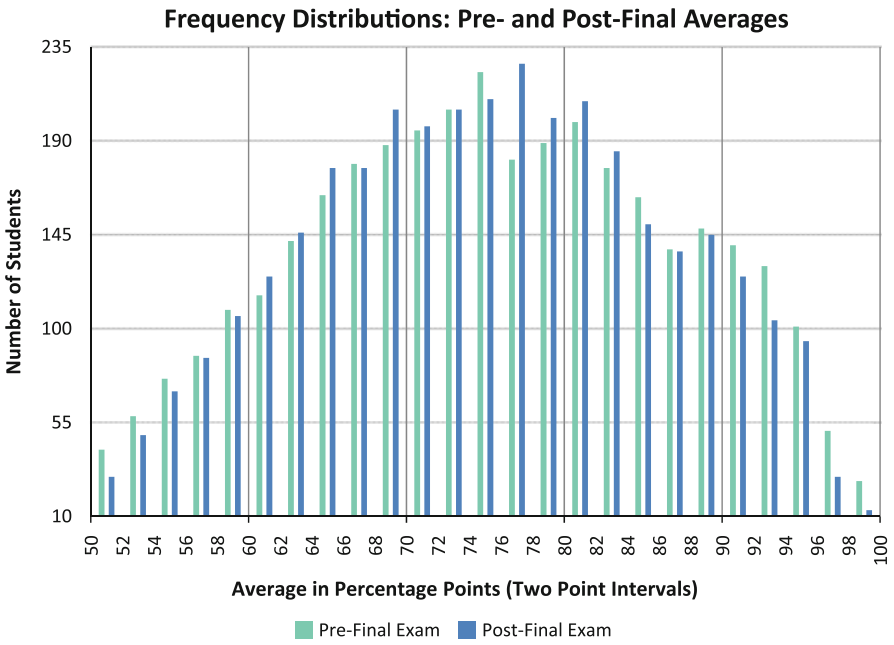
5 Empirical results

We first present preliminary findings, in which the data for Instructors 1–4 are pooled. We then discuss individual instructors' results and present formal hypothesis tests. (Results for Instructor 5 are presented separately in the next section, for reasons explained there.)

5.1 Pooled results: distribution and transition

The top of Fig. 3 contains two frequency distributions of individual course averages, in percent: before taking the final exam, and after. These are grouped into two point intervals: 50.00–51.99, 52.00–53.99, etc. The distribution is approximately normal, with a mean of 75 and a standard deviation exceeding ten: all four grade thresholds are relevant. Threshold incentive effects, if they exist, should be revealed in a systematic bunching of post-exam final averages just above multiples of ten. This does not happen: there is a little bunching above the B/C threshold, but none elsewhere. Many students' averages do change after taking the final exam, up or down, but these tend to offset, so the pre- and post-exam distributions are similar.

These dynamics, and summary evidence on the bunching of final averages, are presented in the transition matrix that comes next in Fig. 3. Each student is classified by the unit digit of their unrounded pre- and post-exam course average: 0 or 1 placing them in the bottom two points of the standard ten-point range, 2–7 placing them in the middle six points of that range, and 8–9 placing them at the top. Pre- to



TRANSITION MATRIX

Post-Final → / Pre-Final ↓	Lower Two Points of Range	Middle Six Points of Range	Upper Two Points of Range	*Row Totals*
Lower Two Points of Range	187 (0.27)	361 (0.52)	*150* *(0.21)*	698 (0.203)
Middle Six Points of Range	**346** **(0.17)**	1380 (0.67)	**349** **(0.17)**	2073 (0.604)
Upper Two Points of Range	*154* *(0.23)*	322 (0.49)	185 (0.28)	661 (0.192)
Column Totals	687 (0.200)	2063 (0.601)	684 (0.199)	3434 (1.000)

Fig. 3 Pooled results (Instructors 1–4). *Top* distribution of pre- and post-exam course averages. *Middle* transition matrix. *Bottom* mean deviation of final exam score from trend

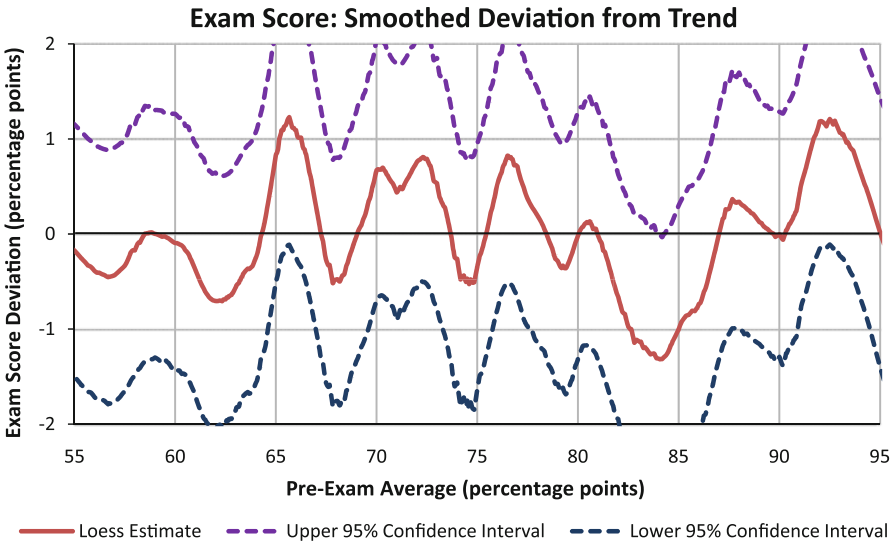


Fig. 3 contiuned

post-exam transition probabilities, along with the total number of students falling in each category, are presented in the interior of the matrix, with row and column totals, and associated proportions, along the outside.

Three pieces of evidence about final exam study behavior are contained in this transition matrix, all versions of the caliper test mentioned above. The first concerns the proportion of students falling in each of the three classifications. Under random placement of students, roughly 20 % should be at the bottom end, 60 % in the middle, and 20 % at the top. If final exam study effort is strategically determined, however, this should not be the case post-exam, with the underlined numbers in the matrix. Instead, the bottom end of each range should have significantly more than 20 % of all students. In fact, the 20–60–20 % distribution is replicated almost perfectly.

The second piece of evidence concerns transitions across classifications after the final exam is taken. If grade incentives matter, these transitions should be asymmetric, with students more likely to move from the upper two points of one grade range to the bottom two points of the next highest range than to go the other way. Thus, the transition probability in the upper-left italicized cell should be smaller than that in the lower-left italicized cell. It is, but only by a small amount, and the difference is not statistically significant.[8]

The last piece of evidence also concerns grade transitions, this time for those students in the second row of the matrix, whose pre-exam average falls in the middle of a grade range. These transitions should also be asymmetric when students are grade-motivated, with movements to the lower two points of a grade range, in the left bolded cell, more frequent than movements to the highest two points, in the right bolded cell. But, in the data, students are as likely to move up as to move down. Final course

[8] A small number of these transitions involve movements in the "right" direction for the "wrong" reasons: for example, a student with a 78.2 average moves down to a 71.4 after a terrible final exam. Accounting for these, the transition probabilities become even more similar.

averages and their pre-/post-exam change exhibit no evidence of threshold-motivated study effort.

5.2 Pooled results: exam scores

Alternatively, this effort could be revealed by unusually strong final exam scores for those students near the thresholds that separate letter grades. Testing for this in the pooled data is complicated by the fact that exam difficulty and the trend relation between the pre-exam average and the final exam score both differ by instructor. Therefore, we first estimated Eq. (1) by instructor (the results can be seen in the Appendix), and calculated the deviation of the exam score from trend, F–$\hat{\text{F}}$, for each student. These deviations were then pooled across all instructors, ordered by the pre-exam average, P, and smoothed with a loess smoother. The results, the pooled equivalent of $\hat{g}(P)$—$\hat{\beta}P - \hat{\gamma}P^2$, approximate the effects of strategic effort on exam scores, and are presented in the bottom graph in Fig. 3.

Strategic, grade-motivated study behavior should generate positive score deviations located slightly below the ten-point thresholds separating letter grades. This is not observed. Estimated deviations in mean exam scores are invariably small (around 1 percentage point), insignificant, and located in the "wrong" places. The peak effort property fails wholly. So do the others. Perturbations in exam scores do not rise faster than they fall, as the sawtooth property predicts, and their slope occasionally drops below negative one, contradicting the stair step property.

5.3 Instructor-specific results

All this evidence suggests the absence of extrinsically, grade-motivated final exam study effort. These findings are confirmed in instructor-specific tests. The estimation results are presented in the Appendix, where a results portfolio for each instructor shows (1) the distribution of final averages, (2) their pre-/post-exam transition, (3) mean final exam performance conditional on the pre-exam average, $\hat{g}(P)$, and (4) the deviation of that performance from trend, $\hat{g}(P)$—$\hat{\beta}P - \hat{\gamma}P^2$. These last two items are the empirical analogs of the bottom and top graphs in Fig. 2, respectively.

Three of these four elements were also shown in Fig. 3, for the pooled data, and the instructor-specific results have the same features. There is virtually no evidence of bunching near the grade thresholds, or of asymmetric transitions in final course averages after taking the final exam; perturbations in exam grades are small, insignificant, and inconsistent with the properties in Section 3. Formal tests of all relevant null hypotheses, five tests for each of four instructors, are found in the center panel of Table 2. Of twenty p values, one is below 0.05, three are below 0.10, and four are below 0.20, roughly as predicted by chance. Even the elusive $p > 0.9$ (DeLong and Lang 1992) is well-represented.

The remaining element, presented third in each results portfolio, shows the mean predicted exam score and, in addition, two other sets of predicted exam scores: the 25th and 75th percentiles. These can be interpreted as estimates for the least grade-motivated and most grade-motivated students (conditional on the pre-exam average).

Table 2 Summary of formal hypothesis tests (p values)

	Caliper test: final averages	Caliper test: transition rates	Nonparam. regression: exam score perturbation	Parametric regression: exam score
Null hypotheses (distribution of test statistic) →	1. The proportion of students in the lower range does not exceed 0.2 (z) 2. The proportions of students in the lower, middle, and upper ranges are 0.2, 0.6, and 0.2 (χ^2)	1. The proportion of students moving from the upper range to the lower range equals that going the other way (z) 2. The proportion of students in the middle range moving to the lower range is no larger than that moving to the upper range (z)	In the terminology of Eqs. (1) and (2), the exam score perturbation, $g(P)—\beta P—\gamma P^2$, is identical to 0 (χ^2)	Coefficients on dummies measuring the distance of the pre-exam average from the threshold are zero 1. Ordinary least squares (F) 2. Least absolute distance–likelihood ratio test (Wald test is similar) (χ^2)
Instructor ↓				
Instructor 1	0.189	0.229	0.730	0.821
	0.778	0.688		0.509
Instructor 2	0.673	0.222	0.738	0.499
	0.939	0.845		0.263
Instructor 3	0.435	0.054	0.086	0.723
	0.997	0.784		0.925
Instructor 4	0.795	0.983	0.269	0.339
	0.569	0.046		0.986
Instructor X (from Oettinger 2002)	Parametric test: $p = 0.01$ at most	–	–	0.122 0.047

If only the latter respond to threshold grade incentives, we should observe positive deviations in exam performance at the 75th percentile, located near each grade threshold, even if we don't observe such deviations in the mean. There is no evidence of this, however. These incentives are impotent across the full range of student motivation.

5.4 Comparison

Table 2 also presents results from a parametric specification introduced by Oettinger (2002), which includes a trinomial in the pre-exam average and four interval dummies for the absolute pre-exam distance from the closest grade threshold, in percentage points, [1,2), [2,3), [3,4), and [4,5), with [0,1) being the omitted category. The joint significance of these dummies is taken to imply the existence of strategic effort. (Our theory implies those individuals above the threshold behave differently from those below it, suggesting an appropriate change in specification, and places

restrictions on the relative coefficient magnitudes. Neither is imposed here, however, to maintain comparability with Oettinger.) For each instructor these dummies are jointly insignificant, reinforcing our nonparametric estimates.

Oettinger estimates this model on grades from a micro principles class and finds that strategic effort exists, on the basis of these joint significance tests and some clustering of students' final averages just above the grade thresholds. Still, threshold effects on final exam performance are modest: 1 percentage point on average and 3 percentage points at most. Oettinger's data, compared to ours, are somewhat more favorable to a positive result: the final exam, 40 % of the course average, counts more than most exams in our data. Such modest effects under stronger grade incentives do not conflict too much with our findings, and suggest that the effect of grade incentives on learning is small under more favorable circumstances and nil under less favorable circumstances.

6 Interpretations

Our null finding is, in effect, reduced form: it tells us that threshold grade incentives do not affect exam performance, but does not tell us why. It is possible that extra study does not improve exam performance, but recent work suggests otherwise (Farkas and Hotchkiss 1989; Eren and Henderson 2008; Stinebrickner and Stinebrickner 2008; DeFraja et al. 2010). Perhaps students simply do not respond to such incentives, for the reasons given by educational psychologists or other reasons. Or perhaps the incentive is functionally effete–too weak or amorphous to be effective. Three supplementary tests, which we now present, undermine this last interpretation.

6.1 Marginality

One possibility is that modest improvements in students' final exam scores do not have a palpable chance of changing their final letter grades. Then the incentive is not meaningful enough to have a significant effect. This explanation can be ruled out. Table 1 reports the number of each instructor's "marginal students," whose final letter grade would change if their final exam score was raised or lowered by five points. On average, fully one-quarter of students fall into this category.

Furthermore, this marginality varies as expected, being greatest for those students whose pre-exam average falls near a grade threshold, at least for Instructors 1, 3, and 4. (We could not determine why Instructor 2 did not satisfy this condition.) Across these three instructors, the fraction of borderline students (whose pre-exam average ended in eight or nine) that were marginal was 27, 28, and 30 %. In contrast, the fraction of non-borderline students (whose pre-exam average ended in four or five) that were marginal was 10, 20, and 14 %. By this accounting, the expected returns to study effort vary between borderline and non-borderline students by about a factor of two.

6.2 Magnitude of the incentive

An alternative explanation is that college grades just aren't that important. We believe this is unlikely. College grades do matter to employers (Chia and Miller 2008; Grant

2007; Wise 1975), and the monetary returns to additional study effort, as implied by prior research, are substantial.

Jones and Jackson (1990) estimates indicate that a one letter-grade improvement in one college course raises wages by .23 %. A very similar estimate obtains by comparing the effect of study effort on GPA, from Stinebrickner and Stinebrickner (2008), with the effect of effort on wages, from Babcock and Marks (2010). For the average college graduate, this improvement is worth more than $100 per year, and even a conservative estimate of net present value would exceed $1,000. The marginality results above indicate that a ten-point increase in the final exam score would increase the probability of getting the higher course grade by about 15 percentage points, so the expected returns to this exam score improvement are at least $150. This would seem to be worth the investment in study effort required to achieve this increase in the exam score.

Furthermore, incentive effects are absent even at the pass/fail border. Because of the large financial and time costs of repeating a class, it is very valuable to receive a D instead of an F. (Each class studied in this paper is required for and dominated by business majors, so, for most students, the unattractive alternative to repeating the class is changing one's major. In each class, also, a grade of D is sufficient for the major, though it must be offset by a higher grade elsewhere.) Yet a careful review of the results portfolios in the Appendix reveals no significant evidence of unexpectedly strong exam performance by students on the cusp of passing the class.

6.3 Awareness

A final possibility is that students simply don't recognize when they are near a grade threshold, though this calculation is easily made given the formulae on the course syllabi. Refuting this possibility requires a falsification test—finding another behavior that does exhibit patterns consistent with threshold incentive effects.

Such a test is possible because our final instructor, Instructor 5, allows students to drop their lowest test, including the final exam. This provides additional leverage: we can analyze the exam-taking decision first, and then the conditional exam score second. These results are presented in an abbreviated results portfolio in Fig. 4. The top graph illustrates the probability of taking the final exam, estimated semiparametrically as before, as a function of the pre-exam course average (with semester dummies and a dummy for missing a previous test as controls). This graph, in contrast to the others, exhibits dramatic variation. It clearly establishes the diminishing marginal value of successively higher grades—moving from an F to a D is valued much more than moving from a B to an A. It also indicates that students think incrementally about the exam-taking decision: within each grade range, exam-taking steadily increases as one approaches the grade threshold. (These thresholds are shifted left by about 3 percentage points, because this instructor rounds up generously. The thresholds are known by students prior to deciding whether to take the final exam.) This is implied by the stair step property, which asserts that exam takers' post-effort passing probabilities continuously increase as the pre-exam average approaches the threshold. Furthermore, for two of

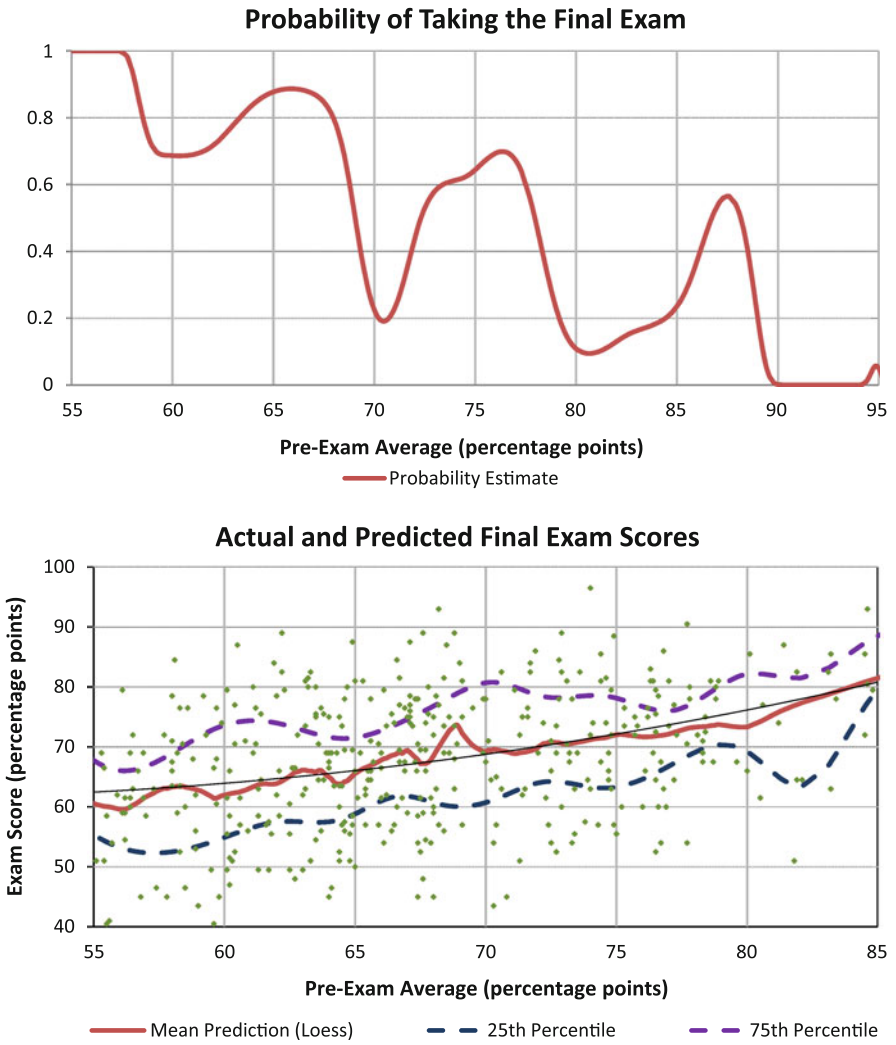


Fig. 4 Abbreviated results portfolio: Instructor 5

the three thresholds in question, the most rapid rise in exam-taking probabilities occurs as one gets reasonably close to the threshold, consistent with the sawtooth property.

The other graph in this figure relates the mean exam score to the pre-exam average for the subset of students that take the final exam (over the limited grade range for which we have sufficient observations), just as in the Appendix. This graph resembles its compatriots–no threshold effect is observed, except perhaps for those just shy of the C/D border. Exam-taking responds to grade incentives, but not exam performance. This indicates that our findings are not explained by students' ignorance of the potential grade benefits of studying harder for the final exam.

7 Conclusions

Though the threshold grade incentives studied here are economically meaningful, they do not inspire the students in our data to strategically raise their exam scores when their grades are most likely to benefit, even when it means the difference between passing and failing. We look to subsequent research to clarify whether this conclusion generalizes to other student populations.

A corollary of our findings is that the A–F grading system has no beneficial incentive or informational properties when criterion-referenced grading is used. (Dubey and Geanakoplos 2010, show that threshold grading can have beneficial properties when norm-referenced grading is used, but argue for the superiority of a criterion-referenced system.) These two properties would tend to go hand in hand. Highly grade-motivated students would tend to cluster just above their preferred threshold, reducing the within-grade spread in performance and making the course grade a good indicator of student achievement in that class. This does not happen in the absence of such motivation. This finding is consistent with our historical review, which found no evidence that threshold grading systems were adopted because they were informationally or motivationally superior. Moving the other direction, to a less "lumpy" plus/minus grading system, improves the informational content of grades, but only marginally (Grant 2007).

Policy-wise, our results sound a note of caution about student incentives. Positive results have been reported from several recent interventions designed to motivate students with substantial, direct monetary rewards, ranging from several hundred to a few thousand dollars (Angrist et al. 2009; Fryer 2010; Jackson 2010). It is probably impractical to implement these too broadly, however, because of the costs involved. Our work indicate that the more modest, yet economically meaningful, indirect rewards provided by grades do not motivate as effectively. In a way, this lack of response is a smaller-scale version of a larger, decades-long problem in the US labor market: the failure, at least among males, for educational achievement to increase despite a massive rise in the returns to schooling (Heckman 2008; Rosin 2010).

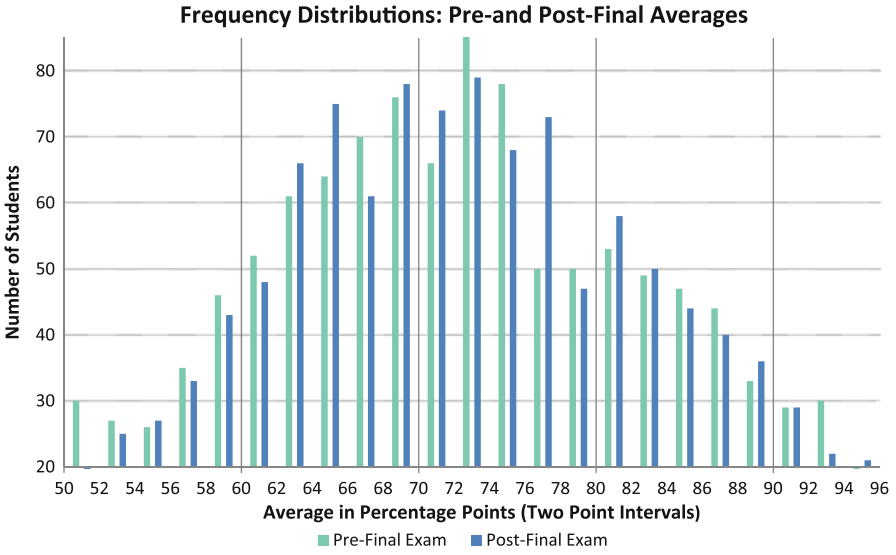
Our results also inform an incipient literature on the design of optimal grading practices (Dubey and Geanakoplos 2010; Zubrickas 2011), by emphasizing a functional distinction between the effects of incentives on algorithmic tasks and on heuristic tasks. This theme, too, is beginning to appear in the literature, in the work of Fryer (2010) discussed above, and in that of Swinton (2010), who finds positive effects of a grading system that was expressly designed to motivate effort both algorithmically and heuristically.

While most policy attention has focused on incentives for teachers or schools, the joint nature of educational production requires that students also be appropriately motivated. This paper's results indicate that designing systems to best accomplish this task represents a formidable challenge for researchers, policymakers, and educators. The magnitude of the incentive, the types of tasks it pertains to, and the presence of intrinsic as well as extrinsic motivation should all be considered when designing such a system.

Acknowledgments We are extremely grateful to the three instructors who generously shared their gradebooks with us: Doug Berg, Natalie Hegwood, and Linda Sweeney. These instructors, along with the authors,

comprise Instructors 1–5 in the text. We also value useful comments received from the associate editor and several referees, Geoffrey Andron, Les Hadsell, and from seminar participants at Baylor University, Sam Houston State University, the Western Economic Association Conference, and the Academy of Economics and Finance. Helpful research assistance was provided by Mohammed Khan, Wade Pate, and Heather Watkins.

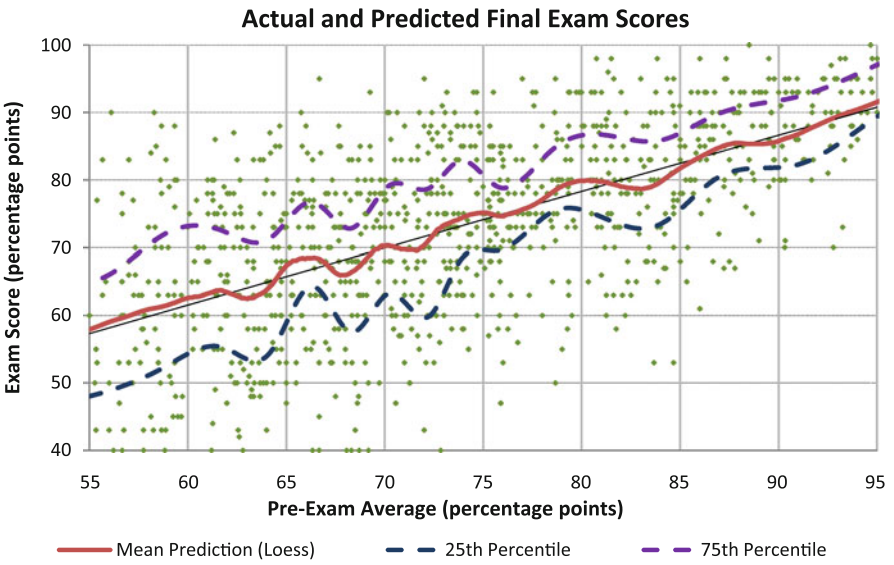
Appendix



TRANSITION MATRIX

Post-Final → / Pre-Final ↓	Lower Two Points of Range	Middle Six Points of Range	Upper Two Points of Range	*Row Totals*
Lower Two Points of Range	73 (0.32)	108 (0.47)	*50 (0.22)*	231 (0.204)
Middle Six Points of Range	**98 (0.14)**	493 (0.71)	**102 (0.15)**	693 (0.612)
Upper Two Points of Range	*52 (0.25)*	94 (0.45)	62 (0.30)	208 (0.184)
Column Totals	223 (0.197)	695 (0.614)	214 (0.189)	1132 (1.000)

Fig. A1 Results portfolio: Instructor 1



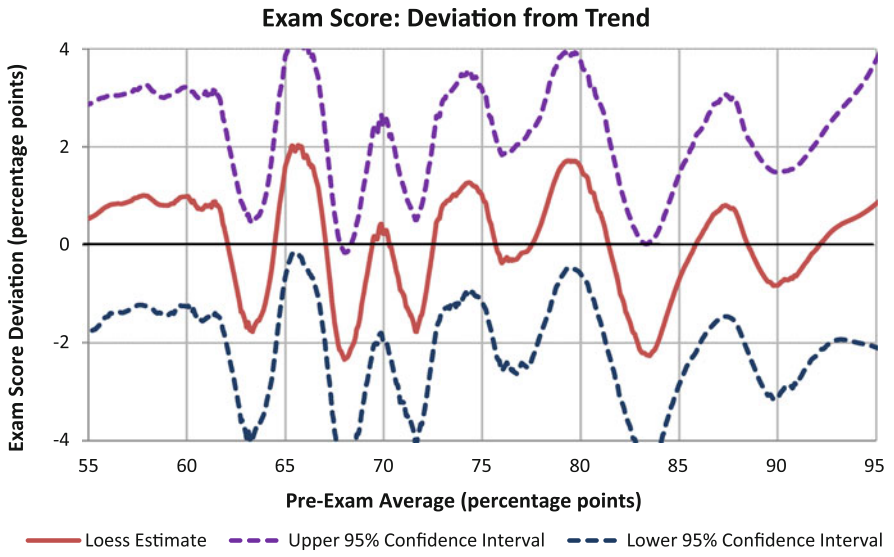
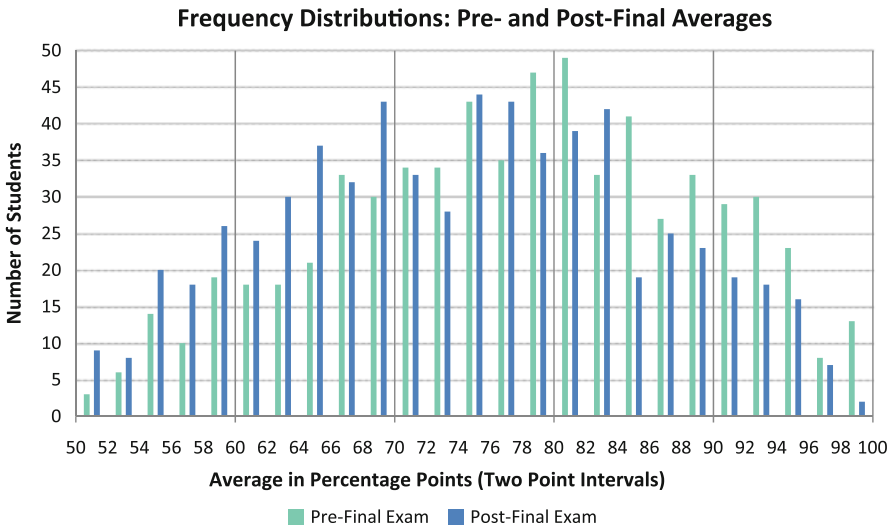


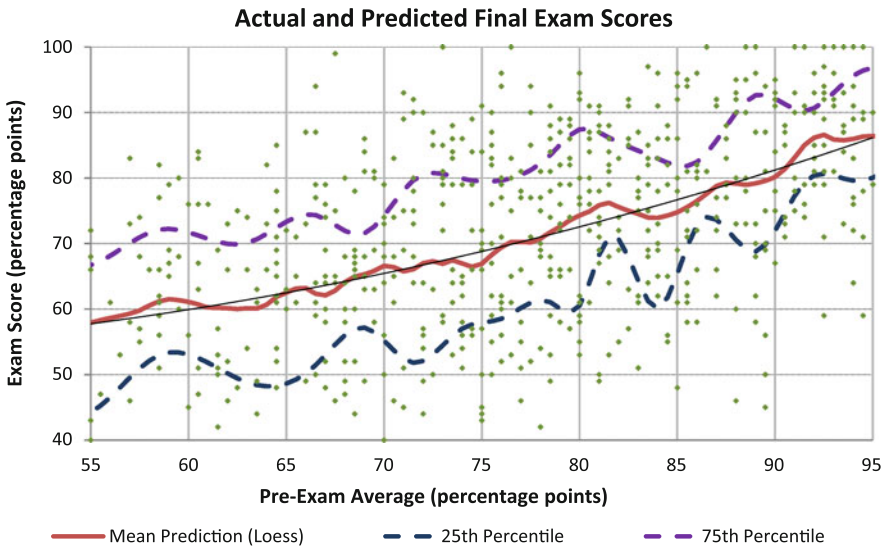
Fig. A1 contiuned



TRANSITION MATRIX

Post-Final → / Pre-Final ↓	Lower Two Points of Range	Middle Six Points of Range	Upper Two Points of Range	*Row Totals*
Lower Two Points of Range	28 (0.20)	77 (0.56)	*32* *(0.23)*	137 (0.21)
Middle Six Points of Range	**58** **(0.15)**	252 (0.67)	**66** **(0.18)**	376 (0.57)
Upper Two Points of Range	*39* *(0.27)*	66 (0.46)	37 (0.26)	142 (0.22)
Column Totals	125 (0.19)	395 (0.60)	135 (0.21)	655 (1.00)

Fig. A2 Results portfolio: Instructor 2



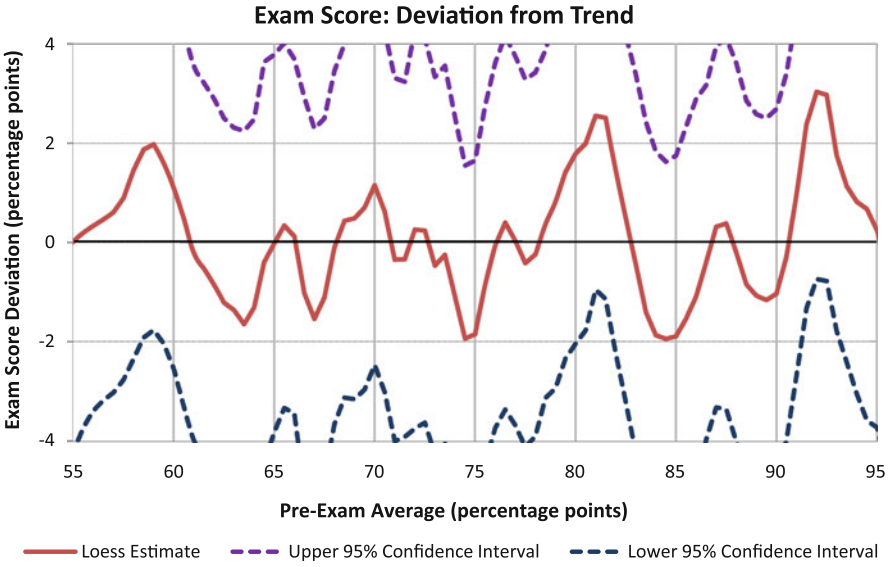
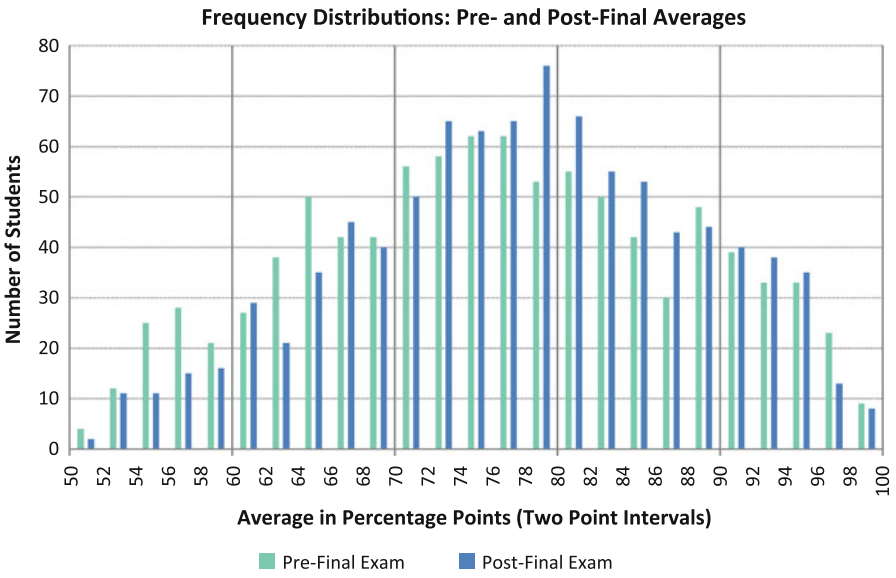


Fig. A2 contiuned



TRANSITION MATRIX

Post-Final → / Pre-Final ↓	Lower Two Points of Range	Middle Six Points of Range	Upper Two Points of Range	*Row Totals*
Lower Two Points of Range	41 (0.23)	109 (0.60)	*32* *(0.18)*	182 (0.19)
Middle Six Points of Range	**103** **(0.18)**	369 (0.63)	**116** **(0.20)**	588 (0.62)
Upper Two Points of Range	*44* *(0.25)*	91 (0.53)	38 (0.22)	173 (0.18)
Column Totals	188 (0.20)	569 (0.60)	186 (0.20)	943 (1.00)

Fig. A3 Results portfolio: Instructor 3

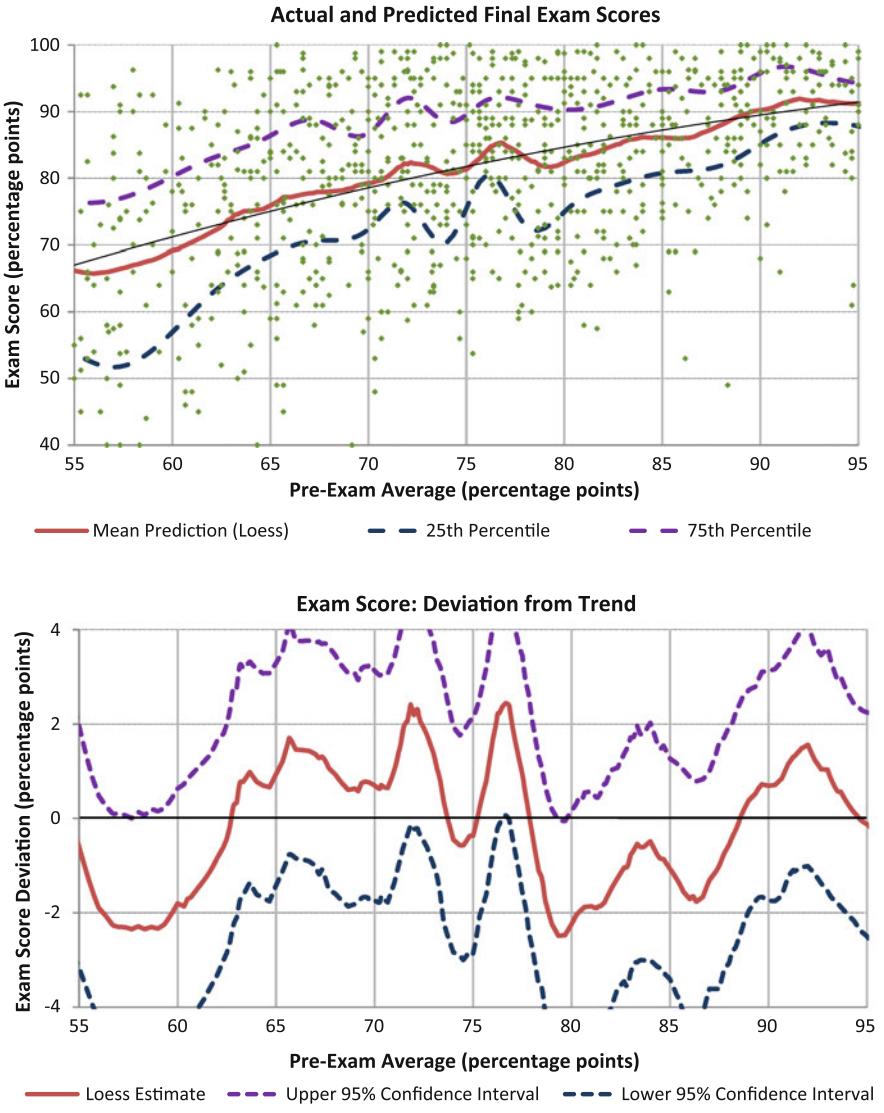
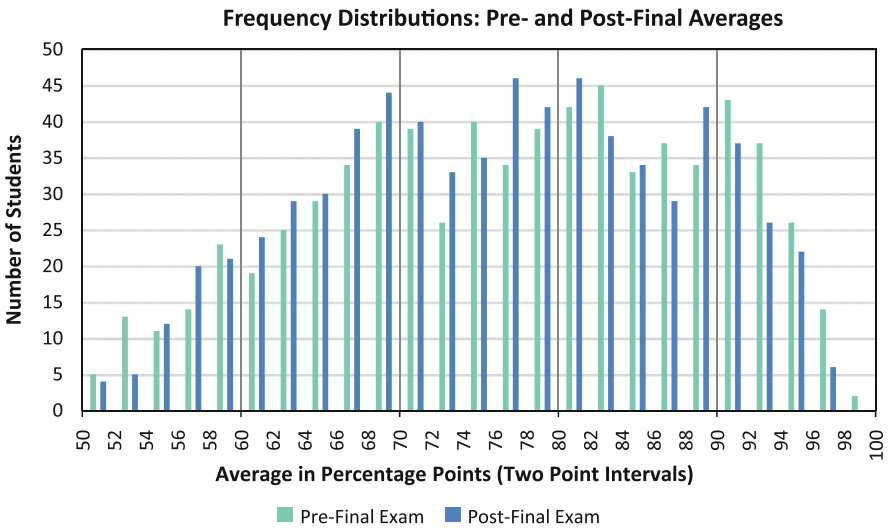


Fig. A3 contiuned



TRANSITION MATRIX

Post-Final → / Pre-Final ↓	Lower Two Points of Range	Middle Six Points of Range	Upper Two Points of Range	*Row Totals*
Lower Two Points of Range	45 (0.30)	67 (0.45)	*36* *(0.24)*	148 (0.21)
Middle Six Points of Range	**87** **(0.21)**	266 (0.64)	**65** **(0.16)**	418 (0.59)
Upper Two Points of Range	*19* *(0.14)*	71 (0.51)	48 (0.35)	138 (0.20)
Column Totals	151 (0.21)	404 (0.57)	149 (0.21)	704 (1.00)

Fig. A4 Results portfolio: Instructor 4

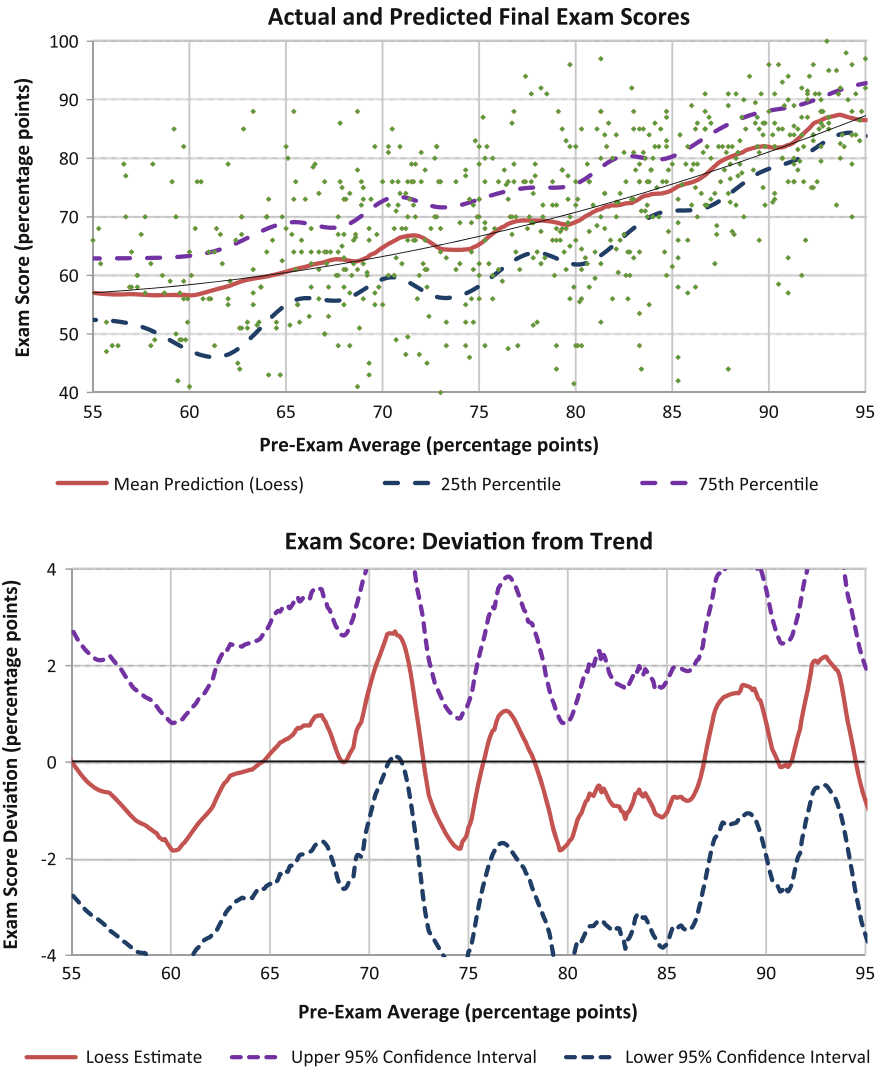


Fig. A4 contiuned

References

Amabile T (1983) The social psychology of creativity: a componential conceptualization. J Pers Soc Psychol 45:357–376

Angrist J, Lang D, Oreopoulos P (2009) Incentives and services for college achievement: evidence from a randomized trial. Am Econ Rev 1:136–163

Ariely D, Gneezy U, Loewentstein G, Mazar N (2009) Large stakes and big mistakes. Rev Econ Stud 76:451–469

Babcock P (2010) Real costs of nominal grade inflation? New evidence from student course evaluations. Econ Inq 48:983–996

Babcock P, Marks M (2010) Leisure college, USA: the decline in student study time. Manuscript, American Enterprise Institute. Florida State University, Tallahassee, FL

Becker W, Rosen S (1992) The learning effect of assessment and evaluation in high school. Econ Educ Rev 11:107–118

Betts J, Grogger J (2003) The impact of grading standards on student achievement, educational attainment, and entry-level earnings. Econ Educ Rev 22(4):343–352

Bishop J (2006) Drinking from the fountain of knowledge: student incentive to study and learn–externalities, information problems and peer pressure. In: Hanushek E, Welch F (eds) Handbook of the economics of education, vol 2. North Holland, Amsterdam

Bonesrønning H (1999) The variation in teachers' grading practices: causes and consequences. Econ Educ Rev 18:89–105

Bonesrønning H (2004) Do the teachers' grading practices affect student achievement? Educat Econ 12: 151–167

Borghesi R (2008) Widespread corruption in sports gambling: fact or fiction?. South Econ J 4:1063–1069

Chia G, Miller P (2008) Tertiary performance, field of study and graduate starting salaries. Austral Econ Rev 41:15–31

Clayson D (2005) Performance overconfidence: metacognitive effects or misplaced student expectations?. J Marketing Educ 27:122–129

DeFraja G, Oliveira T, Zanchi L (2010) Must try harder: evaluating the role of effort in educational attainment. Rev Econ Stat 92(3):577–597

DeLong JB, Lang K (1992) Are all economic hypotheses false? J Polit Econ 6:1257–1272

Dubey Pradeep, Geanakoplos J (2010) Grading exams: 100, 99, 98... or A, B, C? Game Econ Behav 68: 72–94

Elliot A, Zahn I (2008) Motivation. In: Salkind N (ed) Encyclopedia of educational psychology, vol 2. Sage Publications, Thousand Oaks, CA

Eren O, Henderson D (2008) The impact of homework on student achievement. Econom J 11:326–348

Farkas G, Hotchkiss L (1989) Incentives and disincentives for subject matter difficulty and student effort: course grade determinants across the stratification system. Econ Educ Rev 8:121–132

Figlio D, Lucas M (2004) Do high grading standards affect student performance? J Public Econ 88(9,10):1815–1834

Fryer R (2010) Financial incentives and student achievement: evidence from randomized trials. Manuscript, Harvard University, Cambridge, MA

Gerber A, Malhotra N (2008) Publication bias in empirical sociological research: do arbitrary significance levels distort published results? Sociol Method Res 37:3–30

Grant D (2007) Grades as information. Econ Educ Rev 2:201–214

Grant D (2010) The simple economics of thresholds. Manuscript, Sam Houston State University, Huntsville, TX

Grimes P (2002) The overconfident principles of economics student: an examination of metacognitive skill. J Econ Educ 33:15–30

Grove W, Hadsell L (2011) Incentives and student learning. In: Seel N (ed) Encyclopedia of the sciences of learning. Springer, Heidelburg

Grove W, Wasserman T (2006) Incentives and student learning: a natural experiment with economics problem sets. Amer Econ Rev 2:447–452

Hadsell L (2010) Achievement goals, locus of control, and academic success in economics. Am Econ Rev 100:272–276

Heckman J (2008) Schools, skills, and synapses. Econ Inq 46:289–324

Iacus S, Porro G (2008) Teachers' evaluations and students' achievement: how to identify grading standards and measure their effects. Manuscript, University of Trieste, Trieste
Jackson CK (2010) A little now for a lot later: a look at a Texas advanced placement incentive program. J Human Resour 45:591–639
Jones E, Jackson J (1990) College grades and labor market rewards. J Human Resour 2:253–266
Merva M (2003) Grades as incentives: a quantitative assessment with implications for study abroad programs. J Stud Int Educ 2:149–156
Oettinger G (2002) The effect of nonlinear incentives on performance: evidence from "Econ 101". Rev Econ Stat 84:509–517
Pagan A, Ullah A (1999) Nonparametric econometrics. Cambridge University Press, Cambridge
Rosin H (2010) The end of men. The Atlantic 306(1): 56–72
Smallwood M (1935) An historical study of examinations and grading systems in early american universities: a critical study of the original records of Harvard, William and Mary, Yale, Mount Holyoke, and Michigan from their founding to 1900. Harvard University Press, Cambridge
Starch D, Elliott E (1912) Reliability of the grading of high-school work in english. School Rev 20:442–457
Starch D, Elliott E (1913) Reliability of grading work in mathematics. School Rev 21:254–259
Stinebrickner T, Stinebrickner R (2008) The causal effect of studying on academic performance. BE J Econ Anal Poli 8,1:Article 14
Stipek D (1996) Motivation and instruction. In: Berliner D, Calfee R (eds) Handbook of educational psychology. Simon and Schuster, New York
Swinton O (2010) The effect of effort grading on learning. Econ Educ Rev 29:1176–1182
Vendantam S (2008) When play becomes work. Washington Post, July 28, p A2
Wilbrink B (1997) Assessment in historical perspective. Stud Educ Eval 23:31–48
Wise D (1975) Academic achievement and job performance. Am Econ Rev 65:350–366
Yatchew A (1998) Nonparametric regression techniques in economics. J Econ Lit 36:669–721
Zubrickas R (2011) Optimal grading. Manuscript, University of Zurich, Zurich