Sam Houston State University
Department of Economics and International Business
Working Paper Series

THE SIMPLE ECONOMICS OF THRESHOLDS: GRADES AS INCENTIVES

Darren Grant
William B. Green

SHSU Economics & Intl. Business Working Paper No. 09-01
January 2009

Abstract:

This paper examines how grade incentives affect student learning across a variety of ourses at two universities, using for identification the discrete rewards offered by the standard A-F letter grade system. We develop five predictions about effort provision in the presence of the thresholds that separate these discrete rewards, only one of which has been previously tested in the economics literature generally. Surprisingly, all are rejected in our data. Either these grade incentives do not influence student effort appreciably on the margin, or the additional effort is ineffective.

THE SIMPLE ECONOMICS OF THRESHOLDS: GRADES AS INCENTIVES[*]

Darren Grant
Department of Economics and International Business
Sam Houston State University
Huntsville, TX 77341-2118
dgrant@shsu.edu

William B. Green
Department of Economics and International Business
Sam Houston State University
Huntsville, TX 77341-2118
eco_wbg@shsu.edu

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 five predictions about effort provision in the presence of the thresholds that separate these discrete rewards, only one of which has been previously tested in the economics literature generally. Surprisingly, all are rejected in our data. Either these grade incentives do not influence student effort appreciably on the margin, or the additional effort is ineffective.

JEL Codes: I21, A22, D10

Keywords: educational assessment, thresholds, behavioral incentives

*** This paper has eleven pages of figures that are best viewed in color. ***

[*] We are extremely grateful to the three instructors who generously shared their gradebooks with us: Doug Berg, Natalie Hegwood, and Linda Sweeney. We also value helpful comments received from seminar participants at Baylor University and Sam Houston State University, and the research assistance of Mohammed Khan, Wade Pate, and Heather Watkins.

We remain actively interested in obtaining other sets of grades to analyze, especially from upper-division classes and from universities with different profiles than those studied here. Student confidentiality need not be compromised in order to share those grades with us: the only information needed is the pre-exam and post-exam course average, the final exam score, and an identifier for the semester in which the course was taught. Instructors who have taught at least five hundred students in the same course at the same university and are willing to share this data with us are encouraged to contact either author.

In school the efforts of the teacher, the producer of educational services, are augmented by those of the student, the consumer of those services. This special feature of educational production necessitates that the teacher be given the authority to incentivize the student; the substantial labor market value afforded to educational credentials, based on the achievements of each institution's graduates, suggests this authority should be used to generate more effort than the student would privately prefer to give (see also Correa and Gruver, 1987).

Post-secondary education invests this authority almost wholly in grades, which determine whether the student receives credit for the course and signals her level of performance if successful. Yet despite grades' importance, prevalence, and interesting economic features, their incentive properties have received surprisingly little academic attention, as discussed below. With an extensive theoretical and empirical analysis of grades as incentives, this paper tries to fill that gap.

In doing so, we rely on a curious but consequential feature of the typical American grading system: the presence of thresholds, which divide grades into discrete units of A, B, C, D, and F. In the neighborhood of these grade thresholds the marginal benefit of improved performance is highly nonmonotonic, counter to standard economic assumptions. While thresholds such as these are a common feature of economic life, however, their positive and normative properties have not been fully developed. We introduce a basic model of thresholds that generates a sequence of robust predictions that can be applied to a wide range of economic activity, including the study effort induced by grades, and tested elegantly with simple nonparametric methods.

With these 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 indicate that grades are weak incentives: on the margin, either they do not motivate students to study, or the additional study effort is ineffective.

I. The Behavioral Effects of Thresholds.

Public and private entities frequently measure performance on a task of interest. While these measurements commonly use a continuous scale, sometimes the information released to the market, or the administratively determined reward, is binary–linked solely to the passing of a threshold. Economically, this is unusual: the marginal benefit of improved performance is nil until one is about to cross the threshold, after which it is nil again. When measurement is imprecise, so passing the threshold is uncertain (conditional on performance), expected marginal benefits are still non-monotonic, rising and falling rapidly in the neighborhood of the threshold, and thus atypical.

Yet thresholds are often observed, even when a continuous system of measurement and reward appears feasible. Table 1 lists several examples that have been examined in the literature, in labor economics, law and economics, the economics of education, and elsewhere. Thus, a unified discussion of the behavioral and normative properties of thresholds is warranted, along with a concordant, comprehensive estimation strategy. This has not yet been accomplished. We now offer such a development, which builds on the existing literature in several respects:

- The theory is presented assuming imperfect measurement of the agent's performance. Perfect measurement, emphasized in existing theory, is simply a limiting case.
- Five behavioral predictions are established, only one of which has been previously tested.
- Conditions under which thresholds can have desirable normative properties are identified, in contrast to previous work that has emphasized the potential perverse effects of thresholds.
- A more general and revealing econometric strategy is introduced. (Regression discontinuity methods use thresholds differently, to assign individuals to treatment and control groups, and so do not apply here.)

Behavioral Effects in Theory. Let there be a behavioral outcome of interest, t, that is additive in endowed "natural ability," ν, and effort, f, and valued by the market at price $\mathbf{p}$ per unit. When t is measured precisely, each individual's effort is chosen to maximize the difference between the rewards from effort, $\mathbf{p}f$, and its cost, $C(f)$. The solution, $f^* = C'^{-1}(\mathbf{p})$, is efficient as long as the price $\mathbf{p}$ is appropriate (there are no externalities, for example). Continuous, perfect measurement provides ideal information to users and appropriate effort incentives: thresholds are not needed (see Costrell, 1994).

It may be impractical to measure t precisely, however, as measurement exhibits diminishing returns. This is certainly true in education. A typical college course might base grades on two hundred multiple choice questions (over several exams); the standard deviation of the final average of a C student in this class is three percentage points. Reducing it to one percentage point would require increasing assessment time ninefold, utilizing the majority of class time for testing.

Under these circumstances direct performance measurement exhibits the classic signal-extraction problem: variation in the measured outcome is attributable partly to population variation in t and partly to error. Let $T = t + \epsilon$, where ϵ is error in measuring the true outcome, independently and normally distributed. When ν is also normally distributed (throughout the population), the market price of a unit increase in T is $\mathbf{p}\sigma_\nu^2/(\sigma_\nu^2+\sigma_\epsilon^2) < \mathbf{p}$, and each individual underprovides effort.[1] The information provided to the market and the effort elicited by agents can be improved, and under the right circumstances thresholds can do this. *Thresholds can be justified by imperfect information.*

Let the testing agency establish a passing threshold normalized, for simplicity, to 0. Instead of releasing T they simply indicate whether or not $T \geq 0$. The market value of passing the threshold

[1] A technical point: this price supports a symmetric sub-game perfect Nash equilibrium to the N-person "effort game," where each person's effort is optimal given everyone else's choices. As each person provides the same amount of effort, the variance of t ex post equals the variance of ν ex ante.

is $\mathbf{P} = (\bar{t}_{\text{PASSERS}} - \bar{t}_{\text{NONPASSERS}})\mathbf{p}$; the probability of passing the threshold, conditional on effort, is now $\Phi(\nu+f)$, where Φ is the cumulative distribution function for ϵ. The expected marginal returns to effort are bell-shaped, centered around zero. Equating these to the marginal costs of effort can yield multiple solutions for f, which may be minima, local maxima, or global maxima.

Figure 1 illustrates. The horizontal axis indexes t, while the vertical axis indexes costs and benefits. These are expressed in logs, so expected marginal benefits form a parabola, and the marginal cost line corresponds to $C(f) = k \bullet \exp(\gamma f)$, with $\gamma>0$ representing diminishing returns or fatigue in the provision of effort, and k normalized to one. This simple model, with performance linear in ability and effort, normally distributed ability and measurement error, and expected "profit" maximization using this cost function, is analytically tractable and sufficient to substantiate certain claims made below.[2] But the key behavioral predictions that we now deduce are more general, and so are supported only with basic geometric arguments.

Five students, A-E, are represented in Figure 1, their upward sloping marginal cost of effort lines beginning at ν_A-ν_E. For sufficiently low ν, as for student A, marginal costs and marginal benefits do not intersect, so f=0: it is too much work to achieve the higher grade. This continues until the extensive margin is reached, where it is optimal to put forth effort (student B). Here the accumulated surplus, where expected marginal benefits exceed marginal costs, equals the accumulated deficit where the reverse is true. (It does not seem this way in the figure, until one remembers the vertical axis is in logs.) This margin may be reached where $t < 0$, as in the figure; if so effort increases until

[2] For completeness, the analytical solutions for effort (when non-zero) are presented here. Effort under perfect measurement: $f_{\text{PERFECT}} = (1/\gamma)\ln(\mathbf{p})$. Effort under direct but imperfect measurement: $f_{\text{IMPERFECT}} = (1/\gamma)[\ln(\mathbf{p})+\ln(\sigma_\nu^2/(\sigma_\nu^2+\sigma_\epsilon^2))] = f_{\text{PERFECT}} - (1/\gamma)\ln(\sigma_\epsilon^2/\sigma_\nu^2)$. Effort under a threshold: $f_{\text{THRESHOLD}} = -(\gamma\sigma_\epsilon^2 + \nu) + (\gamma^2\sigma_\epsilon^4 + 2\gamma\sigma_\epsilon^2\nu + 2\sigma_\epsilon^2\ln(0.4\mathbf{P}/\gamma\sigma_\epsilon))$^½. Depending on the values of the other parameters, $f_{\text{THRESHOLD}} > f_{\text{IMPERFECT}}$ for all ν, some ν, or no ν.

it reaches its maximum, for student C, at the vertex of the parabola, and declines steadily thereafter (student D) until, at sufficiently high, positive ν, it returns to nil (student E). Those with $0 < \nu < \nu_E$ probably will pass without trying, but assessment is uncertain so they put forth "precautionary" effort to raise their chances. If $t > 0$ at the extensive margin, maximum effort occurs there and declines thereafter; C, the point of maximum effort, falls to the right of the vertex of the parabola.

The resulting $\{\nu,f\}$ and $\{\nu,t\}$ loci, in Figure 2, reveal five positive predictions of the theory.

1. **Peak Effort Property:** *Those individuals far below the threshold ($\nu << 0$) put forth little effort; those near it ($\nu \approx 0$) put forth more; those in between put forth the most.* This is a consequence of the extensive margin, which is itself a consequence of the non-monotonic returns to effort. This property has been noted by several other researchers.

2. **Sawtooth Property:** *Effort rises more quickly than it falls; that is, line BC in Figure 2a rises faster than line CE falls, so that the $\{\nu,f\}$ locus takes a sawtooth shape.* Along with the extensive margin, at which effort increases discretely, this follows geometrically as the upward sloping marginal cost of effort intersects with an inverted parabola.

3. **Peak Proximity Property:** *Those individuals who try the hardest (whose ability is argmax $f(\nu)$) have at least a 50% chance of passing the threshold; that is, line OC in Figure 2a has a slope ≤ -1.* An interior maximum, as in Figure 1, is reached where t=0, so $f = -\nu$ and line OC has a slope of -1. Otherwise maximum effort occurs at the extensive margin, where by the stair step property below $f \geq -\nu$, and line OC is more steeply (negatively) sloped.

These three properties hold whether the measurement is perfect or imperfect. The next two require imperfect measurement to hold, with perfect measurement acting as a limiting case.

4. **Precautionary Effort Property:** *Effort is positive at ν=0.* Error in assessing t motivates precautionary effort to increase the individual's chances of passing. (Under perfect measurement, effort is zero at ν=0.)

5. **Stair Step Property:** *More able individuals always have better outcomes than less able individuals; that is, $\Delta f/\Delta\nu > -1$ and $\Delta t/\Delta\nu > 0$.* Beyond point C, better-endowed individuals work less and still have better outcomes. The $\{\nu,t\}$ locus always slopes upward, but fastest near the extensive margin, like the sloping stair step in Figure 2. (With perfect measurement the step is flat; lower effort fully offsets a higher endowment.)

Estimating Behavioral Effects. Figure 2 is also the departure point for estimating thresholds' behavioral effects, because the available data often permit nonparametric estimation of the $\{\nu, f\}$ or $\{\nu, t\}$ loci directly. Compared to the parametric approach adopted in most previous studies, which utilize one or more dummy variables to check for unusually strong outcomes in some pre-specified ν interval near the threshold, this nonparametric approach requires neither a pre-specified interval nor a pre-specified functional form, thus allowing Properties 1-5 all to be examined. And it completely describes the threshold's incentive effects, with graphs in the form of Figure 2.

From these graphs the validity of the properties listed above, or lack thereof, may be obvious–as it is here. Inference follows a natural progression: first the basic incentive effect, the Peak Effort Property, is formally tested against the null that effort is unrelated to proximity to the threshold–that the demeaned nonparametric estimates of $f(\nu)$ equal zero. This can be done with commercial software. If this null is rejected, the other properties can be tested by identifying the empirical equivalents of point C, point D, line BC, and line CE, and comparing their values, slopes, or relative slopes to those predicted in Properties 2-5. Finally, estimates of the parameters $\{\gamma, \sigma_\epsilon, \mathbf{P}\}$ can be constructed from these values using the generalized method of moments, or via direct structural estimation, and rigorous specification tests conducted.[3]

One can also test for the presence of threshold effects using the ex post distribution of T and pre-test/post-test rates of transition from ν to T. Again no distributional or functional form assumptions are necessary, using what is called "the caliper method" (explicated in Gerber and

[3] The parameter $\mathbf{p}$ can be determined from these estimates and the distribution of T. A cautionary note, however: structural estimation is both complicated and problematic. The location of the extensive margin must be computed numerically, and coefficient estimates may be imprecise, because simulations show that significantly different sets of parameter values can generate similar $\{\nu, f\}$ profiles.

Malhotra, 2008; implemented in economics by Borghesi, 2008, and others; and extended here to transition rates): the empirical density of T in a modest interval just above the threshold should exceed that in an interval of equal size just below the threshold. Also, ν-T transitions should be asymmetrical, with more individuals going from slightly negative ν to slightly positive T than going the other way. The null that the two densities, or two transition rates, are equal is easily tested. Our empirical analysis will present evidence of all these types, all of which supports the same conclusion.

II. Normative Properties of Thresholds.

We now examine three reasons a threshold might be actively preferred to a system of direct measurement. To show that certain normative outcomes are *possible* and determine whether they are *probable* we use simulations, presented in Table 2 and described in the note to the table, that compute various social objectives for various combinations of the parameters $\{\gamma,\sigma_{\epsilon},\mathbf{P}\}$.

Motivating. Effort is underprovided under direct, imprecise performance measurement; its expected returns are attenuated, as some effort is inferred to be noise, instead, in the solution to the signal extraction problem. This effort reduction can be large in relative terms, particularly when the efficient level of effort is small to begin with. This is so in our model, for example. (Using the results and nomenclature in footnote 2, $f_{\text{IMPERFECT}}/f_{\text{PERFECT}} \in [0,1)$, and increases in $\mathbf{p}$.)

Under these circumstances, thresholds can improve efficiency by intensifying the effort of individuals near the threshold. The rewards for passing, $\mathbf{P} = (\bar{t}_{\text{PASSERS}} - \bar{t}_{\text{NONPASSERS}})\mathbf{p}$, are magnified by the divergence in effort between passers and nonpassers and, more subtly, by a positive feedback loop in which the increased effort of passers further increases the rewards for passing, and so on.

Nevertheless, the simulations in Table 2 (and many others not reported here) suggest it is not easy to improve effort efficiency this way. Under direct measurement all individuals provide some effort; under the threshold those far from the extensive margin provide little effort, while others near that margin may overprovide effort. As a practical matter, thresholds seem to increase effort efficiency only when agents are so unmotivated under direct measurement that they hardly try at all.

Signaling. Spence (1973) showed that passing an educational threshold can provide valuable information to employers about workers' underlying aptitudes (ν in our model) even when schooling does not develop human capital. But there was no claim that establishing a threshold is an optimal way to do this, because it is not: direct measurement, even if imperfect, is always superior, because unlike the threshold it does not throw away valuable information on which to condition. If the purpose of schooling is signaling, there is no reason to adopt thresholds.

Performance Measurement. While ν is immutable, t is under the agent's control. Consequently, thresholds can generate more accurate information about performance. Unlike direct measurement, where effort and ability need not be related (as in our model), a threshold system engenders great effort by those low-ν individuals who try to pass, but at most a little precautionary effort by high-ν individuals. The two resulting groups, passers and nonpassers, have disparate *cross-group* outcomes but similar *within-group* outcomes–especially passers, with whom information users are probably most interested. These within-group outcomes can be sufficiently similar that the conditional variance of t is lower than it is under a system of direct performance measurement.

This is confirmed with the Table 2 simulations, which also show that this system works best with higher rewards for passing the threshold (higher **P**), which leads to more effort. Bond ratings, which meet this condition, may have been intended to work this way:

> Credit markets are not continuous; a bond that qualifies, though only by a hair, as investment grade is worth a lot more than one that just fails....There is a huge incentive to get over the line. The challenge to investment banks is to design securities that just meet the rating agencies' tests.... But if the [securities] are too risky, Moody's will object... "Every agency has a model available to bankers that allows them to run the numbers until they get something they like and send it in for a rating" (Lowenstein, 2008).

While the potential for gaming is clear, so to is the potential for within-grade risk to cluster together, enhancing the informational value of a discrete rating system.

We now have two potential theoretical explanations for using thresholds in grading, which both rely on imprecision in performance measurement. When students are not strongly motivated to produce human capital, a threshold can augment effort and thus improve efficiency; when they are strongly motivated, a threshold can improve the accuracy of performance information that is provided to employers or other educational institutions.

Identification of Normative Effects. These normative properties of thresholds can be quantitatively assessed from the joint distribution of $\{\nu, f\}$ when the counterfactual effort under direct measurement is known. If this counterfactual, depicted in Figure 2, is f_0, then the net incentive effects are the simple integral $\int(E(f(\nu)) - f_0)g(\nu)d\nu$, where $g(\nu)$ is the density of ν. Given an estimate of σ_ϵ, the informational properties of thresholds can be assessed as well.

If f_0 is not observed, however, it cannot easily be inferred from $\{\nu, f\}$ alone. This would require identification of the structural model parameters, which is challenging (see footnote 3), and then (for market-determined rewards) solution of the Nash equilibrium in which each person's effort is optimal given others' effort choices. (These choices need not all be identical, as they are depicted in the figure, further complicating matters.) Still, qualitative judgements may sometimes be possible, as they are here, guided when appropriate by the simulation results in Table 2.

III. Incentives and Thresholds in Education.

Historical Evidence on Grade Threshold Adoption. We have identified two potential reasons for adopting a threshold grading system. We now turn to the historical record to see which, if any, of these two reasons resulted in today's widely used letter grade system.

There were no formal educational assessment mechanisms until the end of the 14th century, when Dutch schoolmaster Joan Cele organized a large school. Understaffing necessitated grouping students on the basis of mastery, which required examinations, given twice a year for promotion. These innovations spread throughout Europe over the next two centuries, and were extended during the Industrial Revolution, as the state tried to exercise more control over universities' examination processes in order 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, assessment developed along a similar path. In colonial times, college students were given an oral examination near the end of their studies, which chiefly measured students' ability at rote memorization. But these lenient examinations were mostly just "gestures in public relations" (Rudolph, 1977, p. 145). The first defined scale for differentiating students appeared at Yale in 1785, using four tiers, as in English universities. Coupled with written examinations, more intricate grading systems began to develop. 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, later replaced by a one hundred point scale in order to measure achievement more exactly. 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 19th 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.

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: first, by Cele, to group students into a discrete, homogenous classes to expedite cost-effective instruction; second, by Harvard, to weaken "competition for marks"; and third, motivated by Starch and Eliot, to mask the disparity in instructors' grading standards.

Current Research on Grade Incentives. This history suggests that any beneficial properties of threshold grading systems would be purely incidental. Evidence from educational psychology further suggests that academic achievement may not respond positively to grade incentives.[4]

That literature initially emphasized a model in which behavior responded to "extrinsic" reinforcements, such as grades, and in which these reinforcements could be adjusted in an almost Keynesian way to bring about desired outcomes. Over time, however, 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 a large set of 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

[4] This discussion relies on two recent assessments of the field, Stipek (1996) and Elliot and Ista (2008). In the literature on educational assessment, grade incentives receive even less attention. In 2008 the journal *Studies in Educational Evaluation* had thirty-four volumes. A search of titles, abstracts, and keywords in all articles for the word "incentive" yielded a single match, which was not relevant to the topic of this paper.

"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 ideas are just beginning to creep into economics (Vedantam, 2008), and may help explain the most puzzling question in labor economics today, the weak response of college graduation rates to the increased college wage premium (Altonji, Bharadwaj, and Lange, 2008).

As it stands, though, there is little in the economics literature that explores the effects of grade incentives on student achievement. A few studies (including Grant, 2007, and sources cited therein, and more detailed work by Bonesrønning, 1999, 2004) find that more difficult instructors have better learning outcomes, but this might have more to do with teaching methods than incentives, which are not distinguished empirically. A complementary set of studies explore how study effort affects learning (Farkas and Hotchkiss, 1989; Stinebrickner and Stinebrickner, 2008; DeFraja, Oliveira, and Zanchi, forthcoming; also see Schuman et al., 1985). Results are mixed, because of difficulties measuring study effort and, perhaps, because of the variety of populations studied.

Finally, Oettinger (2002) explores how grade thresholds affect final exam performance for college students, as we do, and concludes that they matter. Both this study and ours "control" for all course characteristics and instructional methods, which are identical across students in the same class. But they cannot distinguish between the amount of incentivized effort and the effectiveness of that effort, so if incentives fail, they cannot isolate why. Oettinger's study and ours differ in five main respects: 1) we emphasize economic significance, while he emphasizes statistical significance; 2) his estimates are parametric, and ours mostly nonparametric; 3) our theoretical development, unlike his, emphasizes the role of uncertainty in passing the threshold, which leads to different empirical predictions and "regression specifications"; 4) he studies stronger students than we do; and

5) his incentives are stronger, because the final exam counts more. Below we compare Oettinger's estimates to ours, and argue there is less dissonance between them than appearances suggest.

In summary, the historical record and the academic literature alike are ambivalent on the effectiveness of grade incentives, and do not indicate that the threshold grading system has valuable normative properties. In the empirical work, accordingly, we adopt the standard null hypothesis that these incentives are ineffectual, and then attempt to marshal enough evidence to reject it.

IV. Data and Implementation.

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 3.

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, on 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.

For each student in each course, the data contain all recorded 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-exam and post-exam course averages, as can the student herself. All courses evaluated students, primarily or wholly, on the basis of two to four midterm exams, one of which could sometimes be dropped, and a final examination that was, except for one instructor, mandatory. 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.

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; six-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 almost a century of combined full-time teaching experience; in their first year in our sample each has at least four 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 uses the standard grading scale, in which 90% is an A, 80% a B, 70% a C, and 60% a D; each occasionally bumps up grades just below the threshold, usually without informing students in advance that they do this. In our data, each instructor 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 our theoretical model performance is a function of innate ability or aptitude, but in the data final exam performance is a function of prior test grades, which reflect a combination of ability and "baseline" effort. The model is flexible enough to handle this. Redefine ν as this combination and f as the "strategic" effort perturbation, positive or negative, in response to threshold incentives (or lack thereof). This model solves as before, and the $\{\nu,f\}$ and $\{\nu,t\}$ graphs take the same shapes. Only the Precautionary Effort Property does not carry through.

This redefinition does not affect estimation of the $\{\nu,t\}$ relation. This is done directly with a semiparametric regression of final exam scores on the pre-exam average and a set of time (semester/year) dummies, added to control for temporal variation in final exam difficulty. Estimation is conducted using a loess smoother, with the smoothing parameter set to cleanly resolve perturbations in mean exam scores as small as two percentage points in width (see footnote 6).

The $\{\nu,f\}$ relation, on the other hand, can only be estimated within an additive constant, as we cannot be sure that the final exam was just as difficult as the tests that preceded it, or (using the results in Section II) that average study effort for the final exam equals that for the previous tests. Therefore, to establish a base from which to identify perturbations in exam performance, we first parametrically regress the exam score on the time dummies, the pre-exam average, and its square. The perturbations are then revealed by nonparametrically regressing these "detrended" residuals on the pre-exam average, again using the loess smoother. Separate regressions are conducted for each instructor; the "sample" analyzed includes all students who took the final exam, had complete pre-exam data, and earned a pre-exam average of at least 50% (sample sizes are in Table 3).

V. Results.

Results for four instructors are presented in Figures 3-6: Professors Berg, Grant, Green, and Hegwood. Each figure contains a portfolio of results for each instructor, illustrating the distribution of final averages, the change in these averages after taking the final exam, final exam performance conditional on the pre-exam average, and the deviation of that performance from trend. We discuss these four instructors' results collectively.

Distribution and Transition. The first graph in each portfolio is a simple frequency distribution 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. In each case the distribution is approximately normal, as would be expected, with a mean between 70 and 80. Strategic behavior should be reflected in a bunching of post-exam final averages just above the ten-point grade thresholds, but this does not generally happen, with a few possibly random exceptions: B's for Profs. Berg and Grant and D's for Prof. Green. 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 each result portfolio. Each student is classified by the unit digit of their unrounded pre-exam 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-exam to post-exam transition probabilities, along with the total number of students falling in each category, are presented in the interior of the transition matrix, with row and column totals, and associated proportions, along the outside.

Each matrix provides three pieces of evidence about strategic final exam study behavior, all versions of the caliper test mentioned above. The first simply concerns the proportion of students falling in each of the three classifications. Under random placement of students, as for example in the pre-exam average, roughly 20% should be at the bottom end, 60% in the middle, and 20% at the top. This does indeed come to pass in all four classes. Strategic exam-taking behavior, however, implies 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. It never does.

The other evidence involves transitions across classifications after the final exam is taken. Strategic behavior should increase the probability of transitioning from the upper two points of one grade range to the bottom two points of the next highest range, and reduce the probability of going the other way. Thus, the transition probabilities in the upper-left italicized cell should exceed those in the lower-left italicized cell. In the data, differences in these transition probabilities are insignificant for two instructors and significant for two others: one, Prof. Green, in the "right" direction and the other, Prof. Hegwood, in the "wrong" direction: a thoroughly split decision.

Transitions for students in the middle of their grade range, in the second row of the matrix, should also be asymmetric when there is strategic behavior: movements to the lower two points of a grade range, in the left bolded cell, should be more frequent than those to the highest two points, in the right bolded cell. This happens for one instructor, Prof. Hegwood, but there are no significant differences for the other three. In summary, for all four classes, final course averages and their pre-exam/post-exam change exhibit almost no evidence of strategic exam-taking behavior.

Exam Scores. The next figure in each portfolio presents exam scores as a function of students' pre-exam averages: the estimated $\{v,t\}$ relation. Its successor–the last figure in the portfolio–presents the mean deviation between the actual scores and those that would be expected if strategic behavior were absent: the estimated $\{v,f\}$ relation.

The top, multi-layered figure begins with a scatterplot of individual exam scores against the pre-exam average. They exhibit great variation, some of which may be due to differences in effort, and the rest due to inter-semester or inter-student differences in exam difficulty, luck in the choice of questions asked, exam-day health, etc. These are difficult to anticipate, justifying our emphasis on uncertain measurement. Added to this scatterplot are three smoothed sets of predicted exam scores: the mean, in the center line, calculated as described above, along with the 25th and 75th percentiles, calculated by applying quantile regression to exam scores that were adjusted for inter-semester differences in exam difficulty.[5] More motivated students reside at the higher percentiles.

The pre-exam average is also affected somewhat by random factors, so there is regression to the mean that brings each line's slope below one. This mean-reversion need not be constant, however, because the contribution of random factors is smaller at high grades and larger at low grades, as can be seen in the exam scatterplots, so a slight convex shape is expected. It is indeed observed for two instructors (Profs. Grant and Hegwood). The long-arc relation between the dependent and independent variables, therefore, requires at least a quadratic–a form that is, in fact,

[5] The coefficients on the time dummies in the mean regression were used to adjust exam scores for the quantile regression. The flexible functional form in the quantile regressions is achieved by representing the pre-exam average as a combination of a series of knots, calculated using transformation regression. The exam score is then regressed on these knots, in a quantile regression, and the curves in the figure are backed out from these estimates. Further details and copies of all programs are available from the first author.

actively preferred, as the the "polynomial wiggle" that could be introduced by a higher-order polynomial might obscure strategic variation in exam scores.

Strategic behavior should be apparent in positive deviations of mean exam scores from this long arc, or trend, located slightly under the ten-point thresholds for each letter grade. This is apparent only rarely: below the B threshold for Prof. Berg and the D threshold for Prof. Grant, and possibly also below the D threshold for Prof. Hegwood. This is true not just at the mean, but also at the 75th percentile, and even where it is apparent the effect is neither large nor significant.

This is demonstrated in the last figure of each portfolio, which depicts (again using the loess smoother) the deviation of the mean exam score from its long-arc quadratic trend, with accompanying upper and lower 95% confidence intervals. In every case, the point estimates rarely exceed two percentage points, are virtually never significant, and are never significant where they should be–just below the grade threshold. The Peak Effort Property fails thoroughly. Its failure obviates the need to check the others, though there is little evidence that they hold either. Perturbations in exam scores do not rise faster than they fall, as the Sawtooth Property predicts; occasionally E(*T*) slopes down, contradicting the Stair Step Property. Again, grade incentives appear to be ineffective.

For completeness, the center panel of Table 4 presents formal tests of all the null hypotheses just discussed. Of twenty p-values, one is significant at the 5% level and another at the 10% level, as predicted by chance. Even the elusive $p > 0.9$ (DeLong and Lang, 1992) is well-represented.

Comparison. Table 4 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 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. These dummies capture *f* in reverse–effort relative to those on the borderline–and their joint significance implies the existence of strategic effort. (Our theory implies those above the threshold behave differently from those below it, suggesting the specification should be modified accordingly. We do not do this here, however, to maintain comparability with Oettinger.) For each instructor these dummies are jointly insignificant, reinforcing our nonparametric estimates, as do several other robustness tests.[6]

Oettinger estimates this model on grades from a micro principles class at the University of Texas's flagship Austin campus and finds that strategic effort exists, on the basis of these joint significance tests and evidence that students' final averages cluster just above the grade thresholds. Still, even here, threshold effects on final exam performance are modest: one percentage point on average and three percentage points at most. Oettinger's data, compared to ours, are perhaps less representative of American higher education, but more favorable to a positive result: the final exam, 40% of the course average, counts more, and the students he studies are more capable. Such modest effects 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.

Exam Taking. The final instructor for which we have data, Prof. Sweeney, allows students to drop

[6] A variety of nonparametric estimators and smoothing values were feasible. The differences between them can be summarized as follows. First, the choice of estimator is inconsequential. Estimates were also constructed with transformation regression, least absolute difference regression, and nonparametric spline estimators, all to little effect. Second, by choosing the smoothing parameter to resolve perturbations in the nonparametric estimates that are small in width, estimates of strategic effort are *exaggerated*. Using the statistically preferred smoothing value, deviations of effort from trend rarely exceeded one-half percentage point, and were always insignificant. This smoothing value is utilized for the hypothesis tests presented in Table 4.

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 Figure 7. The top graph illustrates the probability of taking the final exam, estimated semiparametrically using transformation regression, as a function of the pre-exam course average (with time dummies and a dummy for missing a previous test as controls). This graph, in contrast to those preceding it, 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 three percentage points, because this instructor rounds up generously. The thresholds are known by students prior to 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 the three thresholds in question, the most rapid rise in exam-taking probabilities occurs as one gets within a reasonable range of the threshold, consistent with the Sawtooth Property.

The other graph in this portfolio 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). This graph resembles its compatriots–no threshold effect is observed, except perhaps for those just shy of the C/D border. Exam-taking appears to respond to grade incentives, but not exam performance. Overall, there is little evidence that students strategically raise their exam scores via increased study effort when their grades are most likely to benefit, even when it means the difference between passing and failing.

VI. Conclusions.

In our data grades either do not motivate on the margin, or the additional effort is ineffective. In neither case would grade thresholds have beneficial incentive properties. Nor, then, do individual course grades provide good information about students' performance. This conclusion is linked to the previous one. Highly grade-motivated students would tend to cluster just above their preferred threshold, making the course grade a good indicator of student achievement in that class. This will not happen in the absence of such motivation. This explains Grant's (2007) quixotic finding that the primary component of grades in micro principles classes at a non-Texas university is not teacher expectations or student ability but unrelated, potentially random factors.

We cannot isolate the root cause of our main finding, but there are only a few possibilities: extrinsic incentives are inherently ineffective, for the reasons given by educational psychologists; the incentives provided by grades are weak, because of imprecision in the measurement of performance or low rewards to passing grade thresholds (but see Grant, 2007, for evidence that college grades matter to employers); or study effort is ineffective. It is also unclear whether our findings generalize to other student populations. To facilitate study of this question, we will share our estimation programs with interested parties–or will execute the estimations on data provided to us.

What is clear is that further study is warranted. Teacher-focused or school-focused educational reforms will be less successful, and less cost-effective, if students are not appropriately motivated. Our study suggests this may not be easy to do.

REFERENCES

Altonji, Joseph, Prashant Bharadwaj, and Fabian Lange. "Changes in the Characteristics of American Youth: Implications for Adult Outcomes," NBER Working Paper 13883 (2008).

Bonesrønning, Hans. "The Variation in Teachers' Grading Practices: Causes and Consequences," *Economics of Education Review*, 18:89-105 (1999).

----. "Do the Teachers' Grading Practices Affect Student Achievement?" *Education Economics*, 12:151-167 (2004).

Borghesi, Richard. "Widespread Corruption in Sports Gambling: Fact or Fiction?" *Southern Economic Journal*, 74, 4:1063-1069 (2008).

Correa, Hector, and Gene Gruver. "Teacher-Student Interaction: A Game Theoretic Extension of the Economic Theory of Education," *Mathematical Social Sciences*, 13:19-47 (1987).

Courty, Pascal, and Gerald Marschke. "An Empirical Investigation of Gaming Responses to Explicit Performance Incentives," *Journal of Labor Economics*, 22, 1:23-56 (2004).

Costrell, Robert. "A Simple Model of Educational Standards," *American Economic Review*, 84, 4:956-971 (1994).

DeFraja, Gianni, Tania Oliveira, and Luisa Zanchi. "Must Try Harder: Evaluating the Role of Effort in Educational Attainment," *Review of Economics and Statistics*, forthcoming.

DeLong, J. Bradford, and Kevin Lang. "Are All Economic Hypotheses False?" *Journal of Political Economy*, 100, 6:1257-72 (1992).

Elliot, Andrew, and Ista Zahn. "Motivation," in N. Salkind, ed., *Encyclopedia of Educational Psychology*, Vol. 2. Thousand Oaks, CA: Sage Publications (2008).

Farkas, George, and Lawrence Hotchkiss. "Incentives and Disincentives for Subject Matter Difficulty and Student Effort: Course Grade Determinants across the Stratification System," *Economics of Education Review*, 8:121-132 (1989).

Friedman, David, and William Sjostrom. "Hanged for a Sheep–The Economics of Marginal Deterrence," *Journal of Legal Studies*, 22, 2:345-66 (1993).

Gerber, Alan, and Neil Malhotra. "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?" *Sociological Methods Research,* 37:3-30 (2008).

Grant, Darren. “Dead on Arrival: Zero Tolerance Laws Don’t Work,” *Economic Inquiry*, forthcoming.

----. “Grades as Information,” *Economics of Education Review*, 26, 2:201-214 (2007).

Healy, Paul. “The Effect of Bonus Schemes on Accounting Decisions,” *Journal of Accounting and Economics*, 7:85-107 (1985).

Iyengar, Radha. “I Would Rather Be Hanged for a Sheep Than a Lamb: The Unintended Consequences of California Three-Strikes Law,” NBER Working Paper 13784 (2008).

Lowenstein, Roger. “Triple-A Failure,” *New York Times Magazine*, April 27, 2008:36.

McEwan, Patrick, and Lucrecia Saltibañez. “Teacher Incentives and Student Achievement: Evidence from a Mexican Reform,” manuscript (2005).

Muradian, Roldan. “Ecological Thresholds: A Survey,” *Ecological Economics*, 38:7-24 (2001).

Oettinger, Gerald. “The Effect Of Nonlinear Incentives On Performance: Evidence From "Econ 101,” *Review of Economics and Statistics*, 84:509-517 (2002).

Perrings, Charles, and David Pearce. “Threshold Effects and Incentives for the Conservation of Biodiversity,” *Environmental and Resource Economics*, 4:13-28 (1994).

Reback, Randall. “Teaching to the Rating: School Accountability and the Distribution of Student Achievement,” *Journal of Public Economics*, 92:1394-1415 (2008).

Rudolph, F. *Curriculum. A history of the American undergraduate course of study since 1636*. San Francisco: Jossey Bass (1977).

Schuman, Howard, Edward Walsh, Camille Olson, and Barbara Etheridge. “Effort and Reward: The Assumption that College Grades Are Affected by the Quantity of Study,” *Social Forces*, 63:945-966 (1985).

Smallwood, Mary L. *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*. Cambridge: Harvard University Press (1935).

Spence, A. Michael. “Job Market Signaling,” *Quarterly Journal of Economics*, 87:355-374 (1973).

Starch, Daniel, and Edward Elliott. “Reliability of the Grading of High-School Work in English,” *The School Review*, 20:442-457 (1912).

Starch, Daniel, and Edward Elliott. “Reliability of Grading Work in Mathematics,” *The School Review*, 21:254-259 (1913).

Stinebrickner, Todd, and Ralph Stinebrickner. "The Causal Effect of Studying on Academic Performance," Working Paper, University of Western Ontario (2008).

Stipek, Deborah. "Motivation and Instruction," in D. Berliner and R. Calfee, eds., *Handbook of Educational Psychology*. New York: Simon and Schuster (1996).

Tufte, Edward. *Beautiful Evidence*. Cheshire, Connecticut: Graphics Press (2006).

Vendantam, Shankar. "When Play Becomes Work," *Washington Post*, July 28, 2008, p. A2.

Wilbrink, Ben. "Assessment in Historical Perspective," *Studies in Educational Evaluation*, 23:31-48 (1997).

Figure 1. Analysis of the Effort Decision, Conditional on Ability.

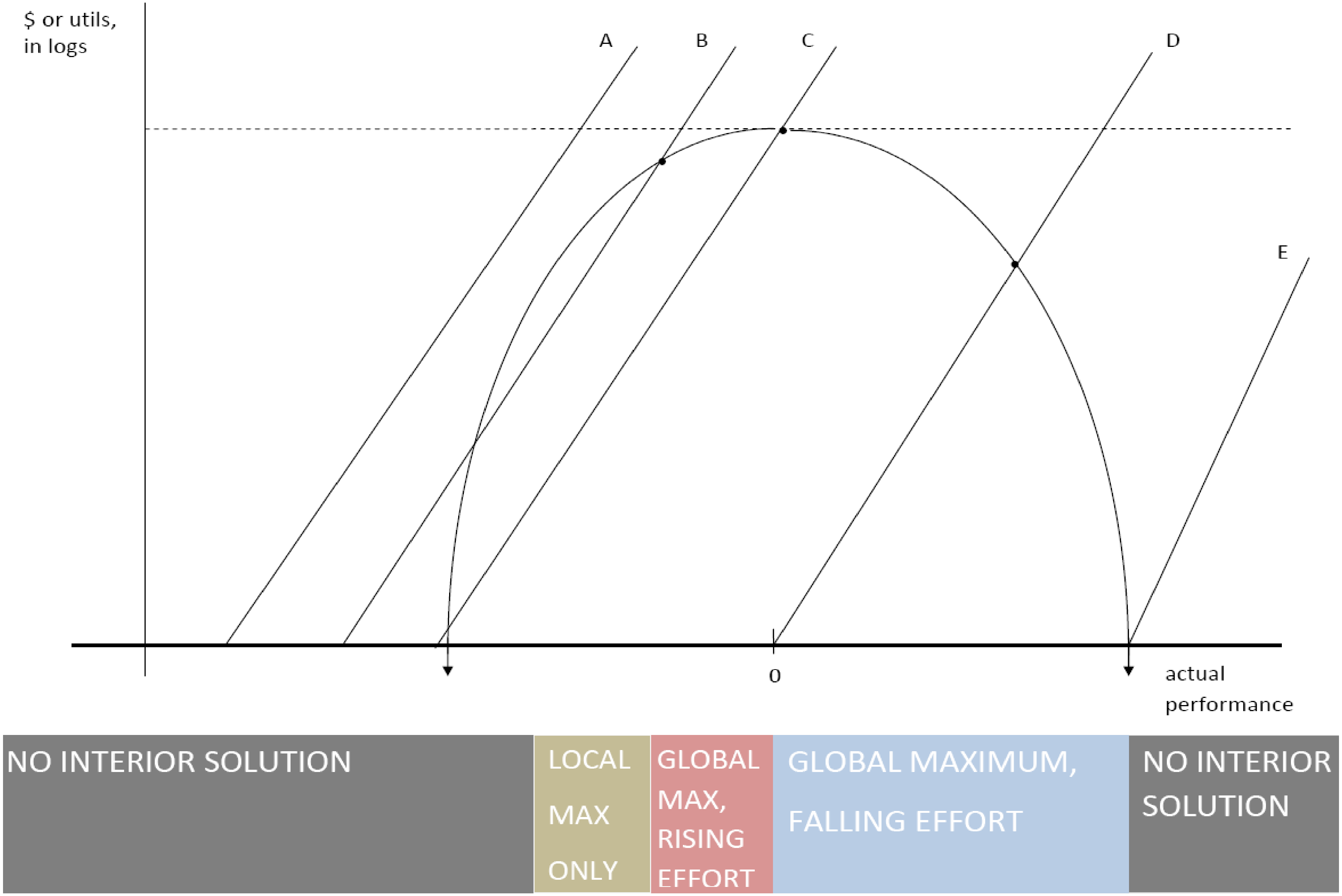


Figure 2. Top: Ability-Effort Locus. Bottom: Ability-Performance Locus.

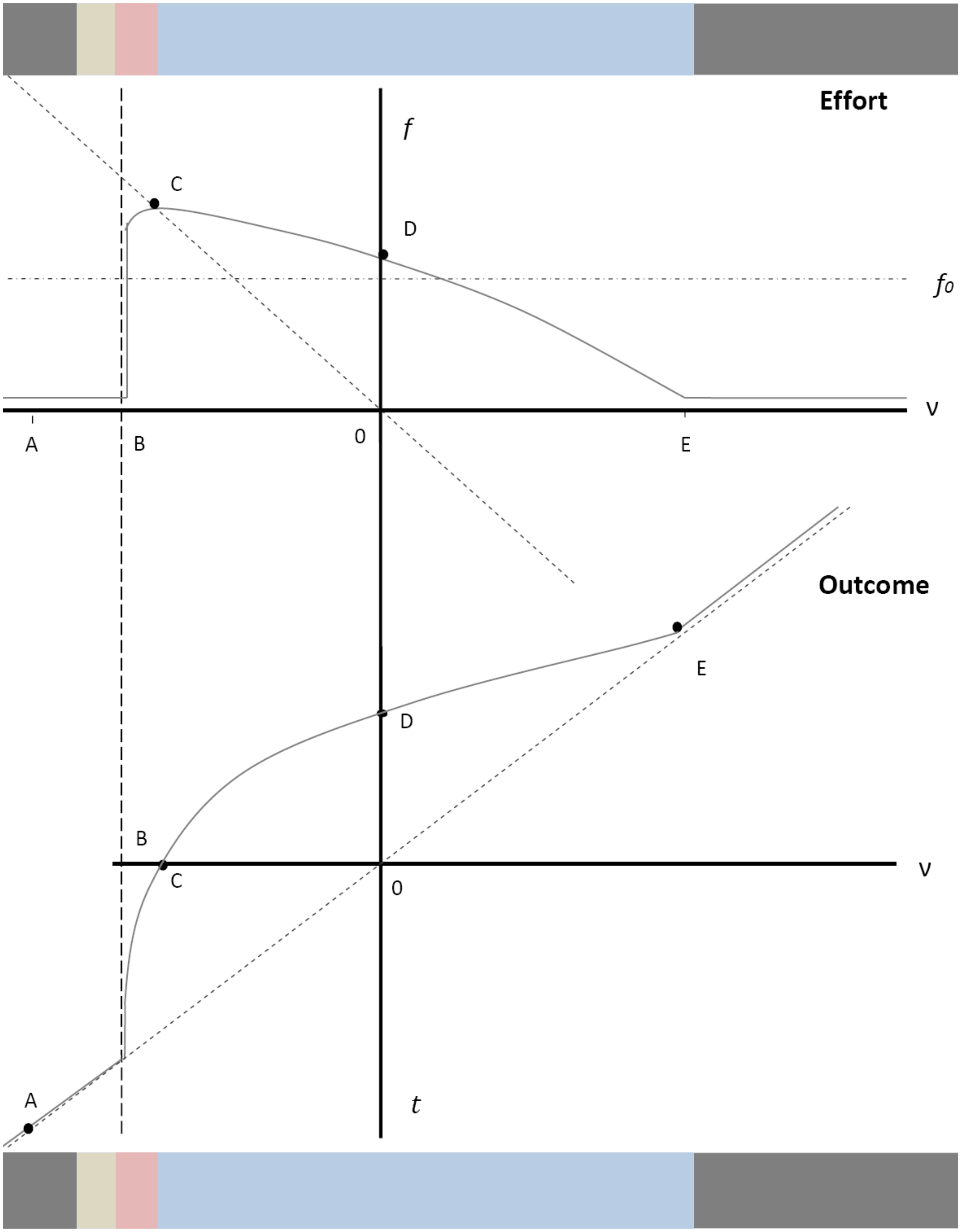
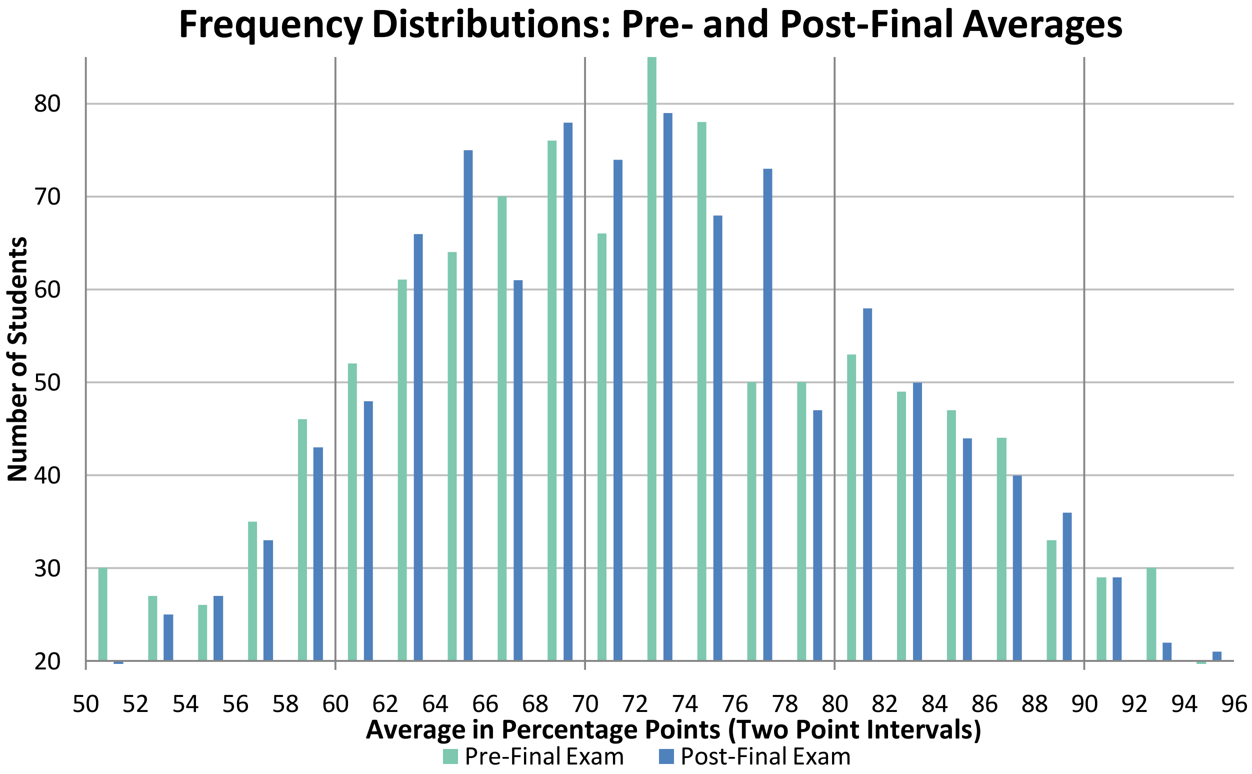
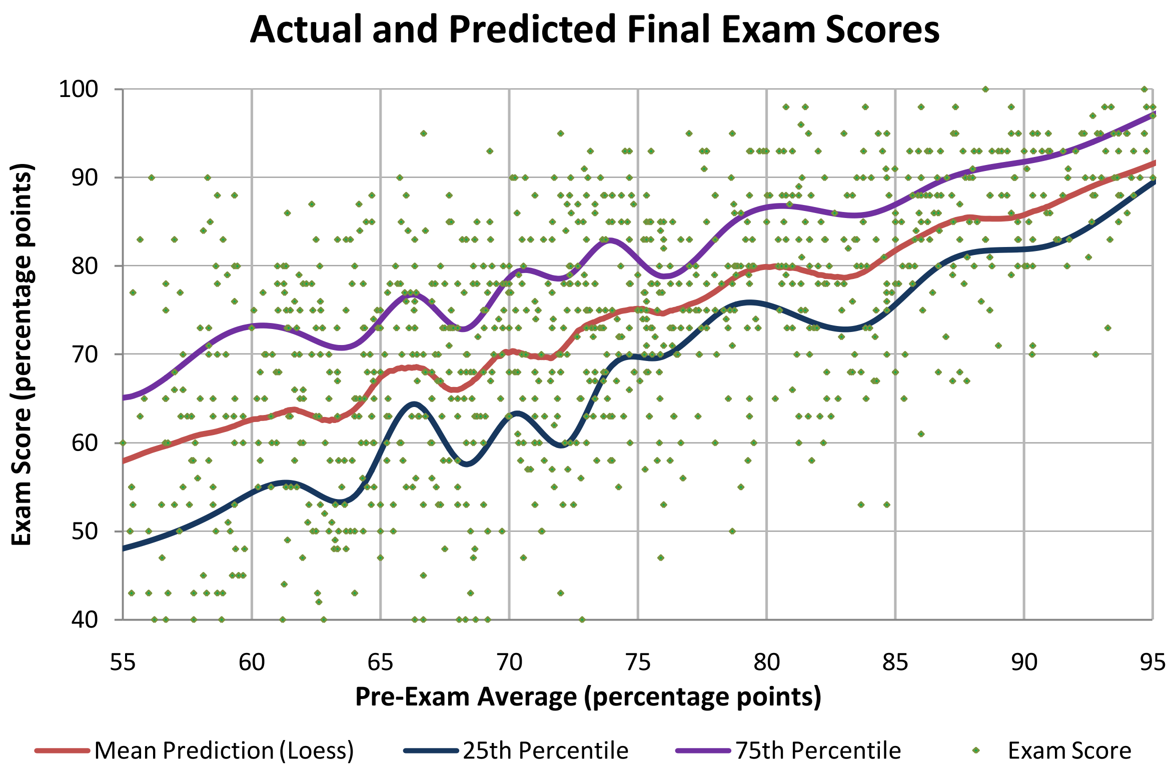


Figure 3. Results Portfolio: Berg.



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)


Actual and Predicted Final Exam Scores
100
90
80
70
60
50
40
Exam Score (percentage points)
55
60
65
70
75
80
85
90
95
Pre-Exam Average (percentage points)
Mean Prediction (Loess)
25th Percentile
75th Percentile
Exam Score

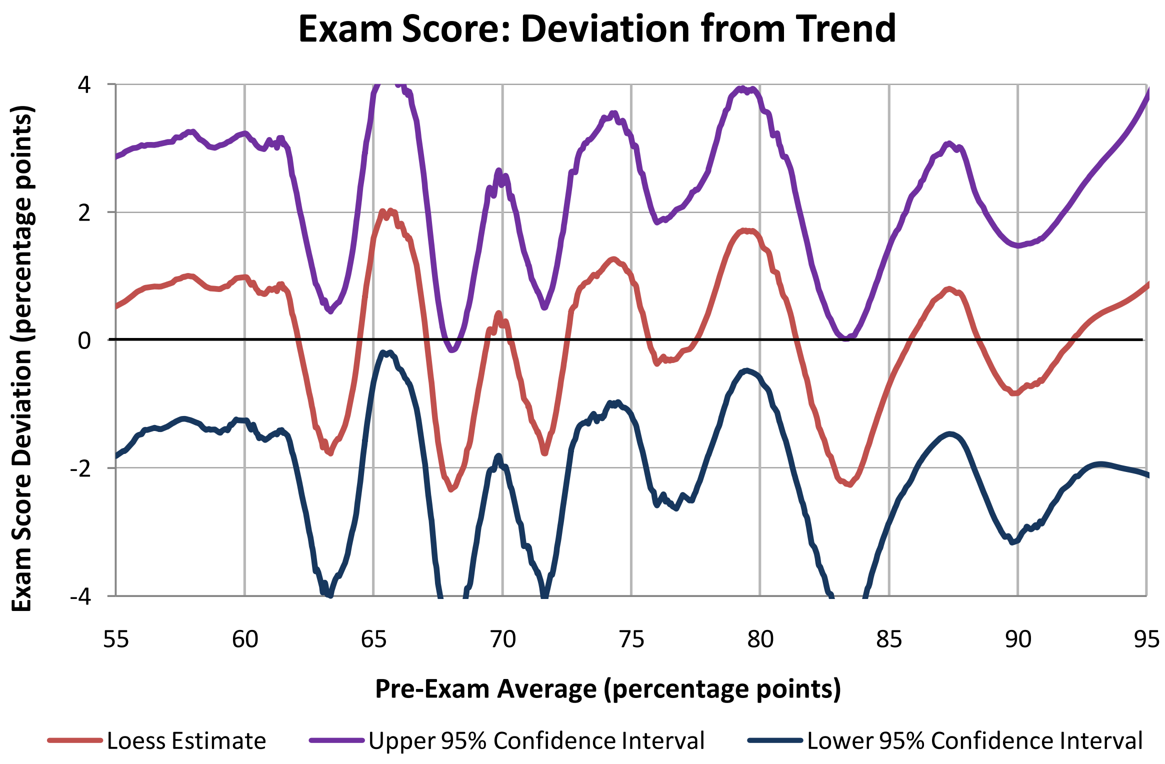
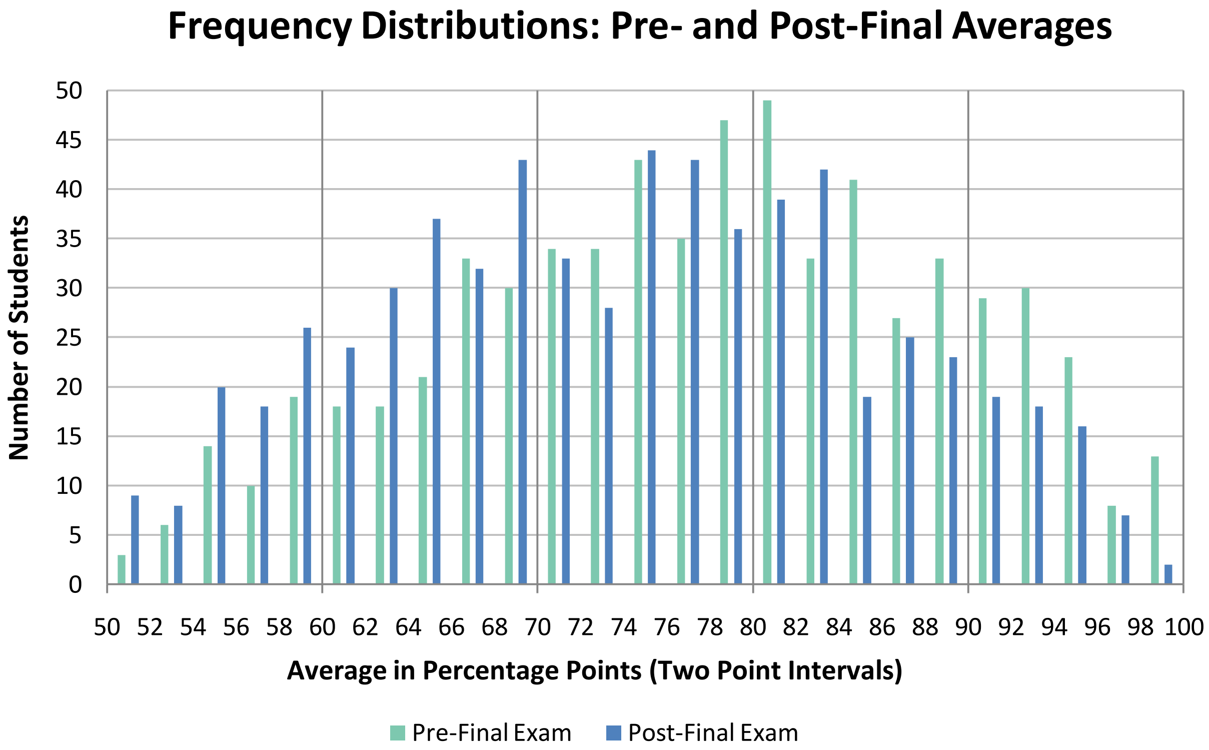
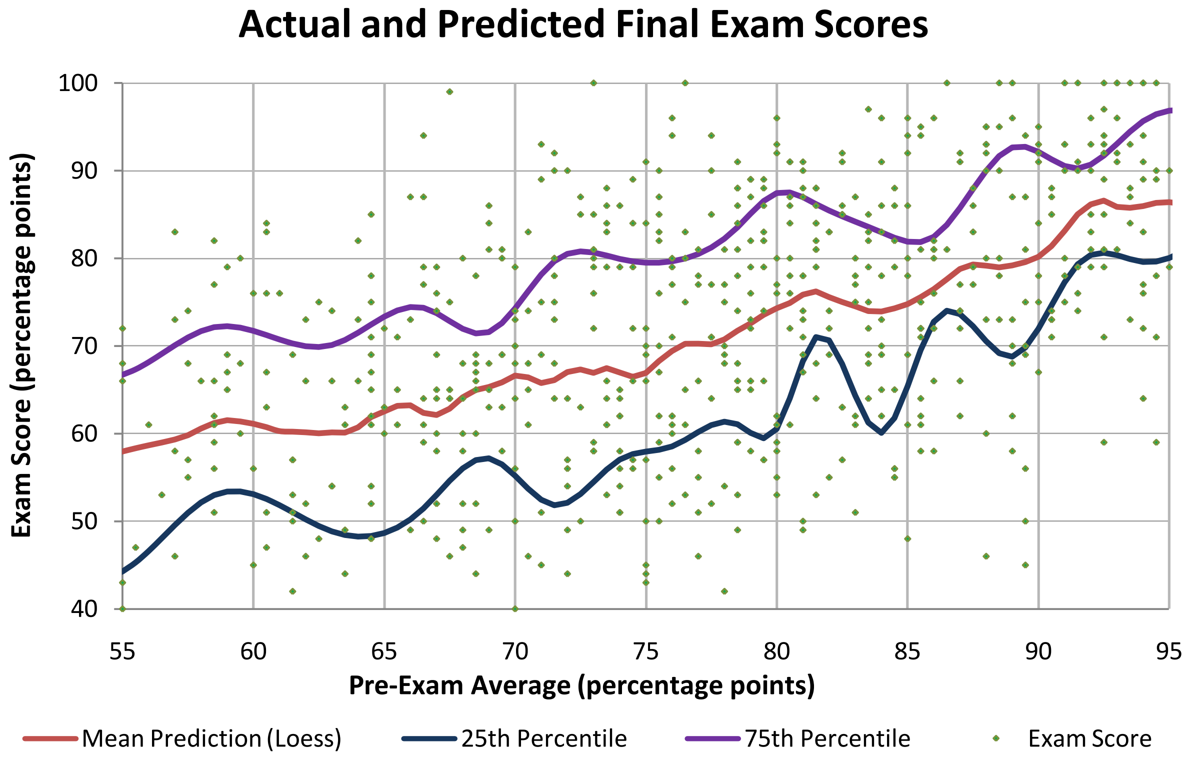

Exam Score: Deviation from Trend
4
2
0
-2
-4
Exam Score Deviation (percentage points)
55
60
65
70
75
80
85
90
95
Pre-Exam Average (percentage points)
Loess Estimate
Upper 95% Confidence Interval
Lower 95% Confidence Interval

Figure 4. Results Portfolio: Grant.



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)


Actual and Predicted Final Exam Scores
Exam Score (percentage points)
100
90
80
70
60
50
40
55
60
65
70
75
80
85
90
95
Pre-Exam Average (percentage points)
Mean Prediction (Loess)
25th Percentile
75th Percentile
Exam Score

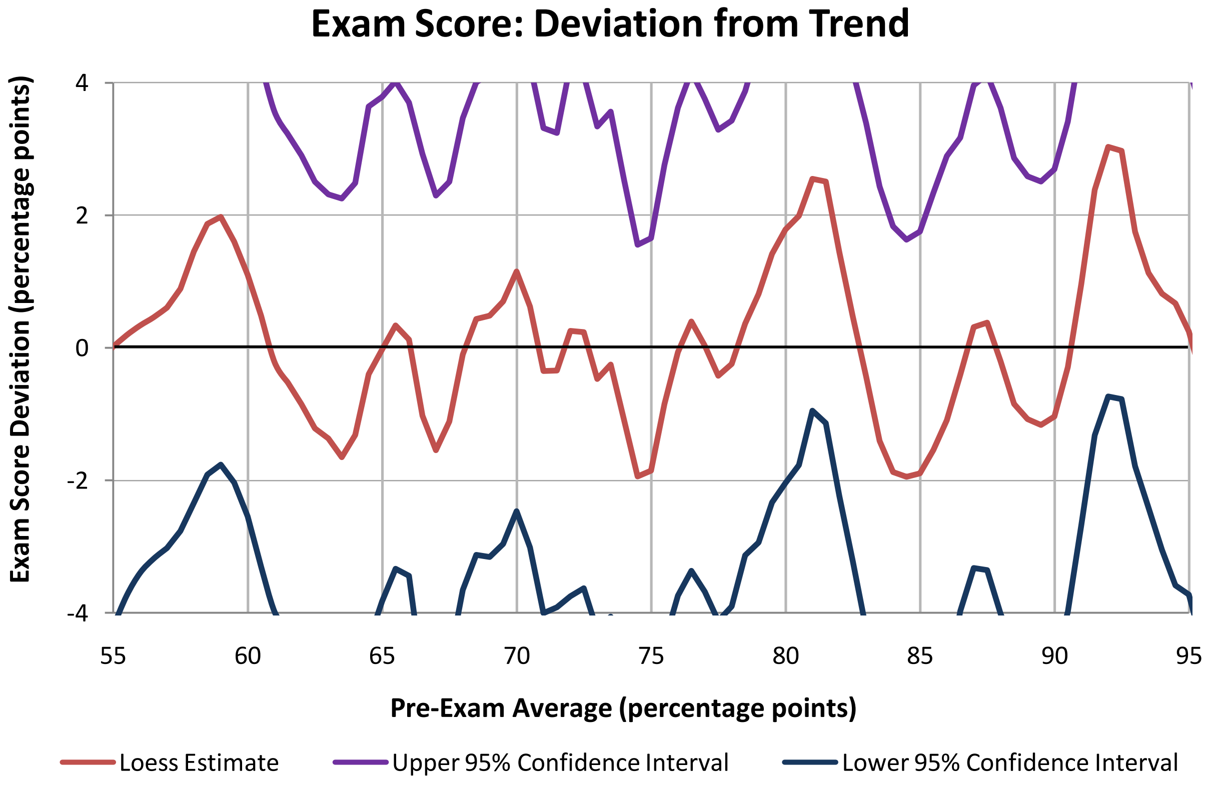
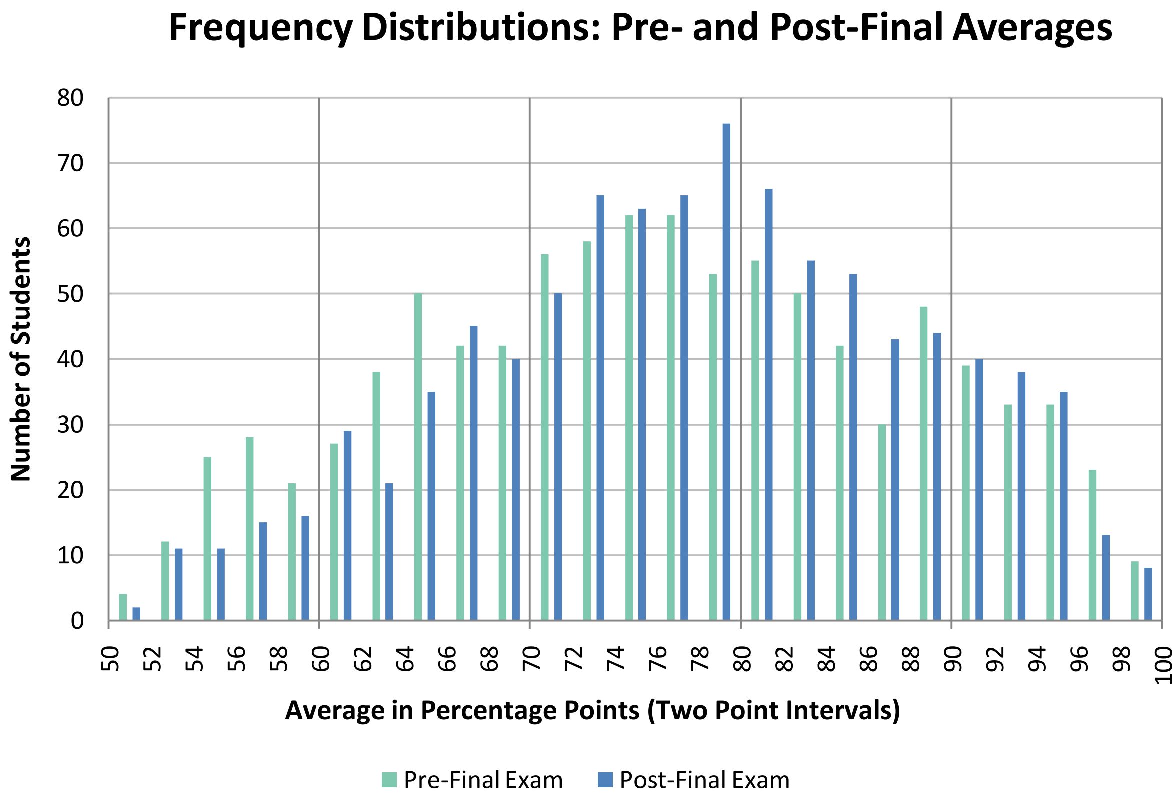
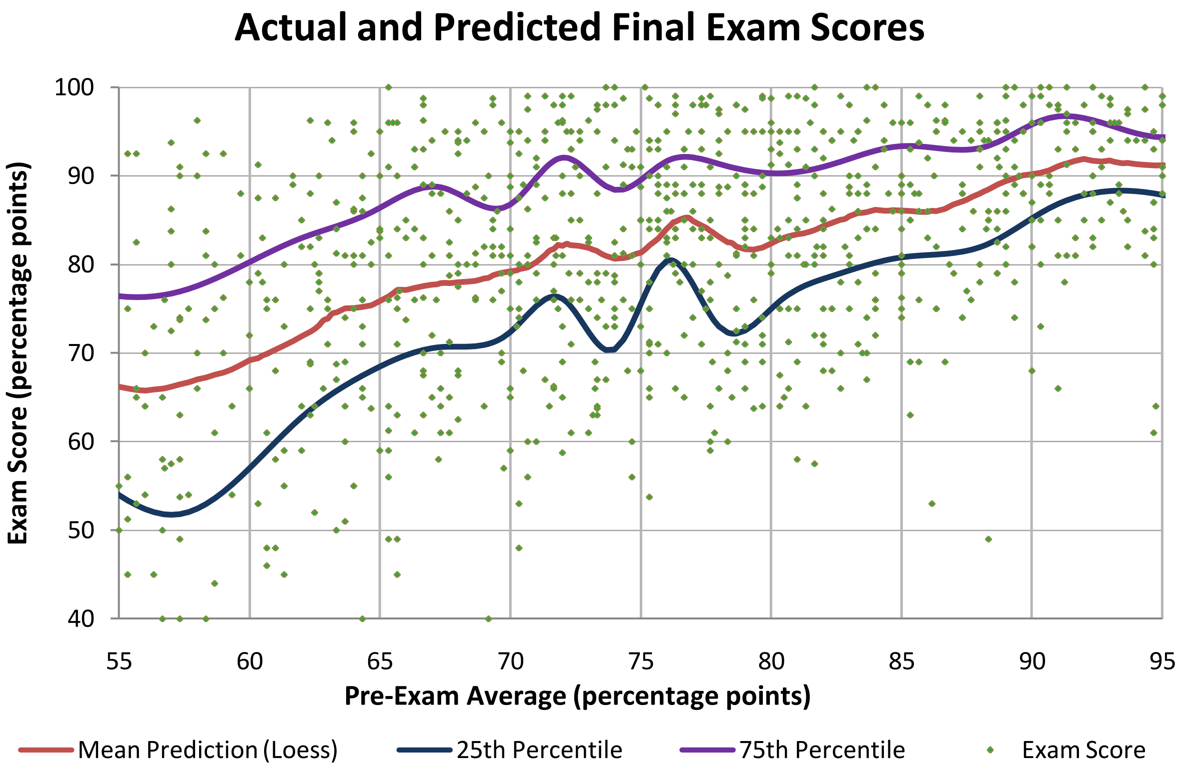

Exam Score: Deviation from Trend
Exam Score Deviation (percentage points)
4
2
0
-2
-4
55
60
65
70
75
80
85
90
95
Pre-Exam Average (percentage points)
Loess Estimate
Upper 95% Confidence Interval
Lower 95% Confidence Interval

Figure 5. Results Portfolio: Green.



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)


Actual and Predicted Final Exam Scores
Exam Score (percentage points)
100
90
80
70
60
50
40
55
60
65
70
75
80
85
90
95
Pre-Exam Average (percentage points)
Mean Prediction (Loess)
25th Percentile
75th Percentile
Exam Score

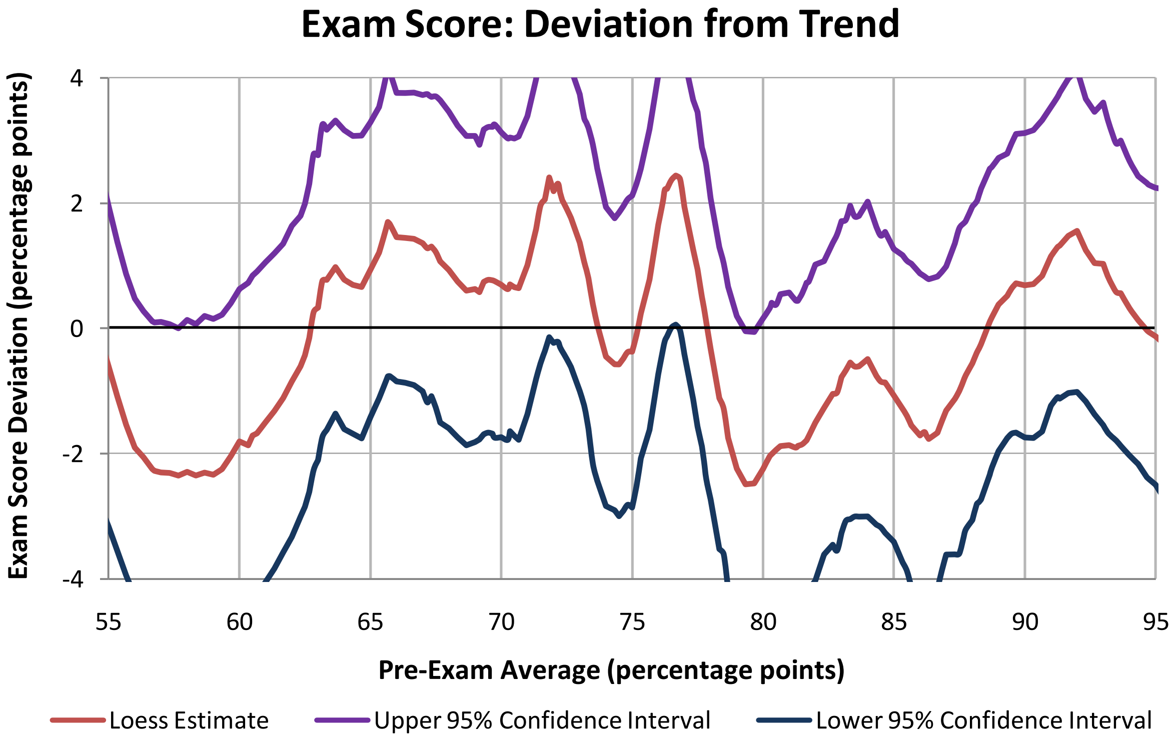
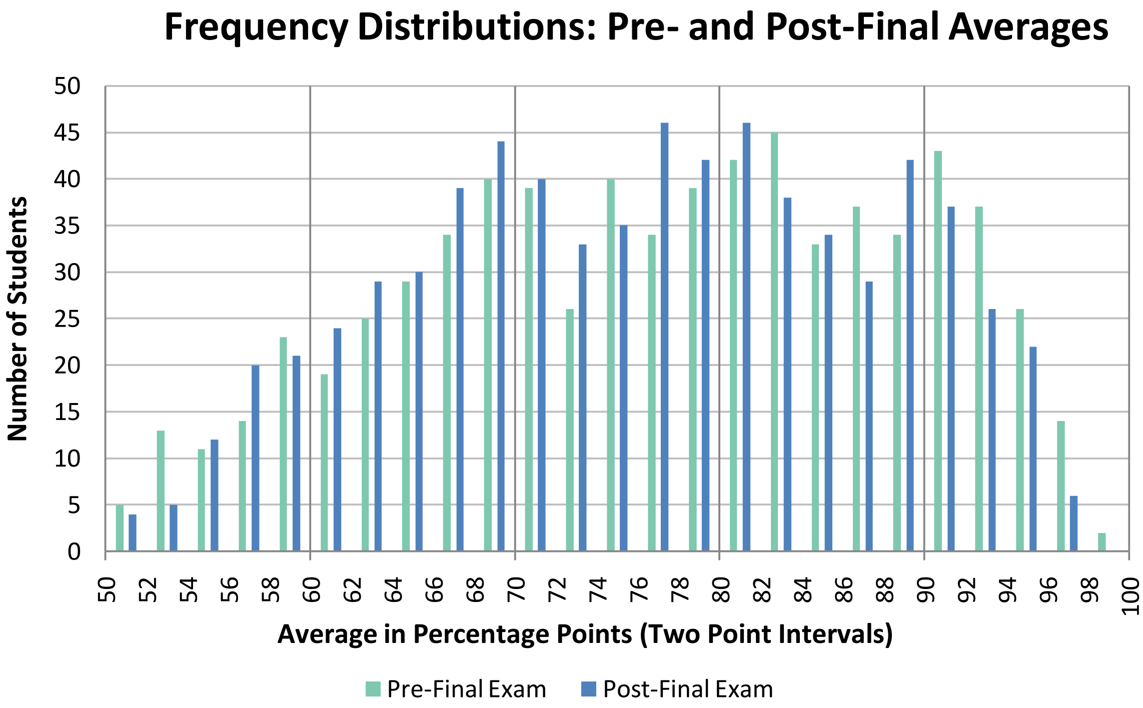
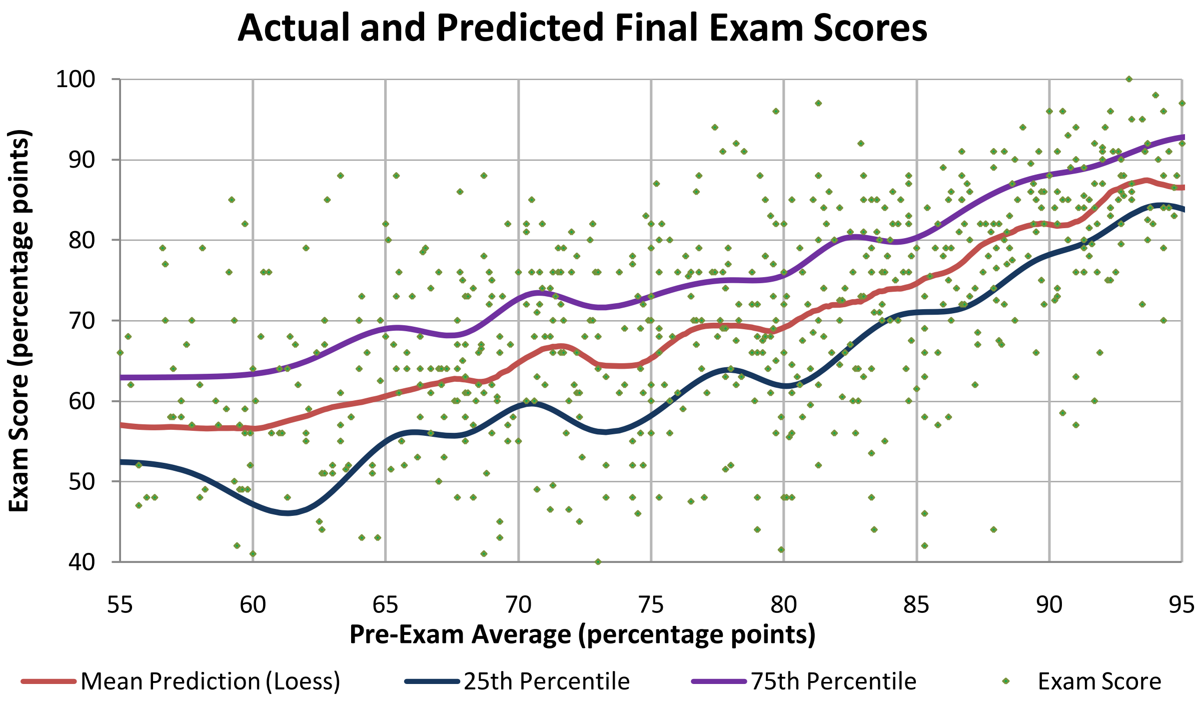

Exam Score: Deviation from Trend
Exam Score Deviation (percentage points)
4
2
0
-2
-4
55
60
65
70
75
80
85
90
95
Pre-Exam Average (percentage points)
Loess Estimate
Upper 95% Confidence Interval
Lower 95% Confidence Interval

Figure 6. Results Portfolio: Hegwood.



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)


Actual and Predicted Final Exam Scores
Exam Score (percentage points)
100
90
80
70
60
50
40
55
60
65
70
75
80
85
90
95
Pre-Exam Average (percentage points)
Mean Prediction (Loess)
25th Percentile
75th Percentile
Exam Score

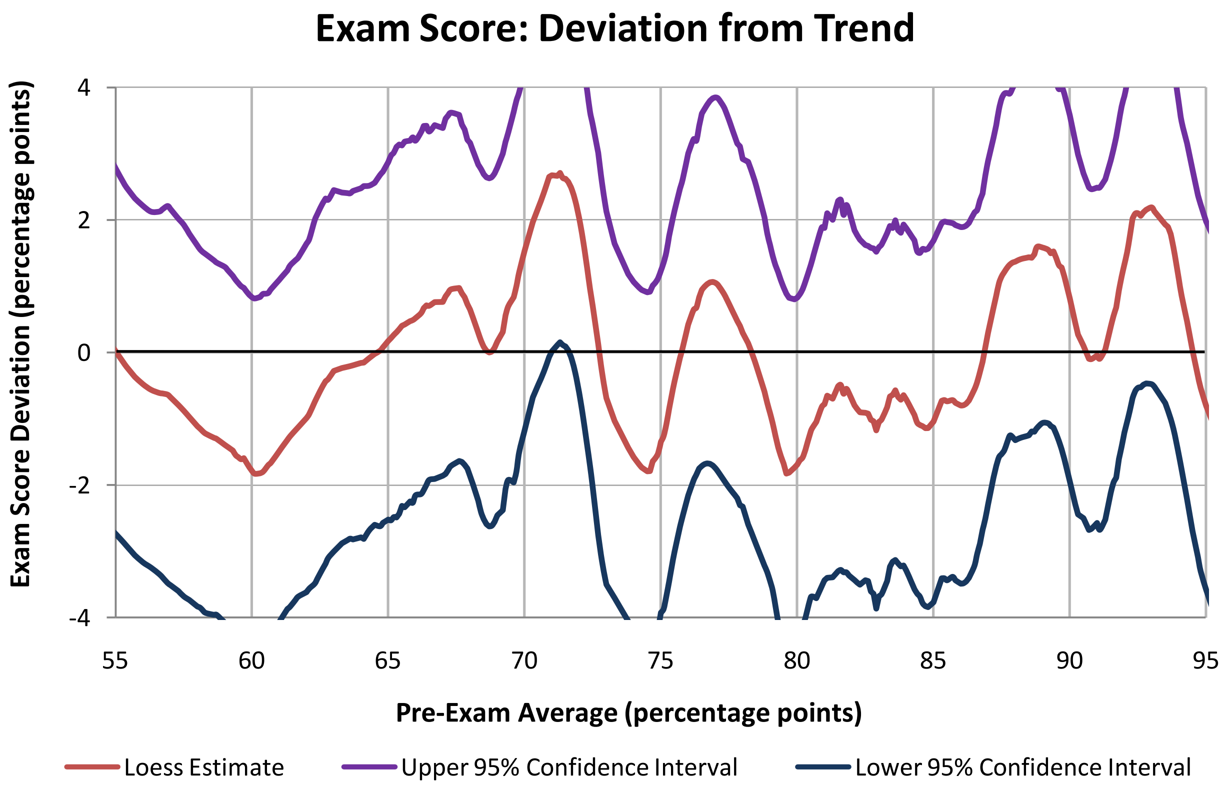
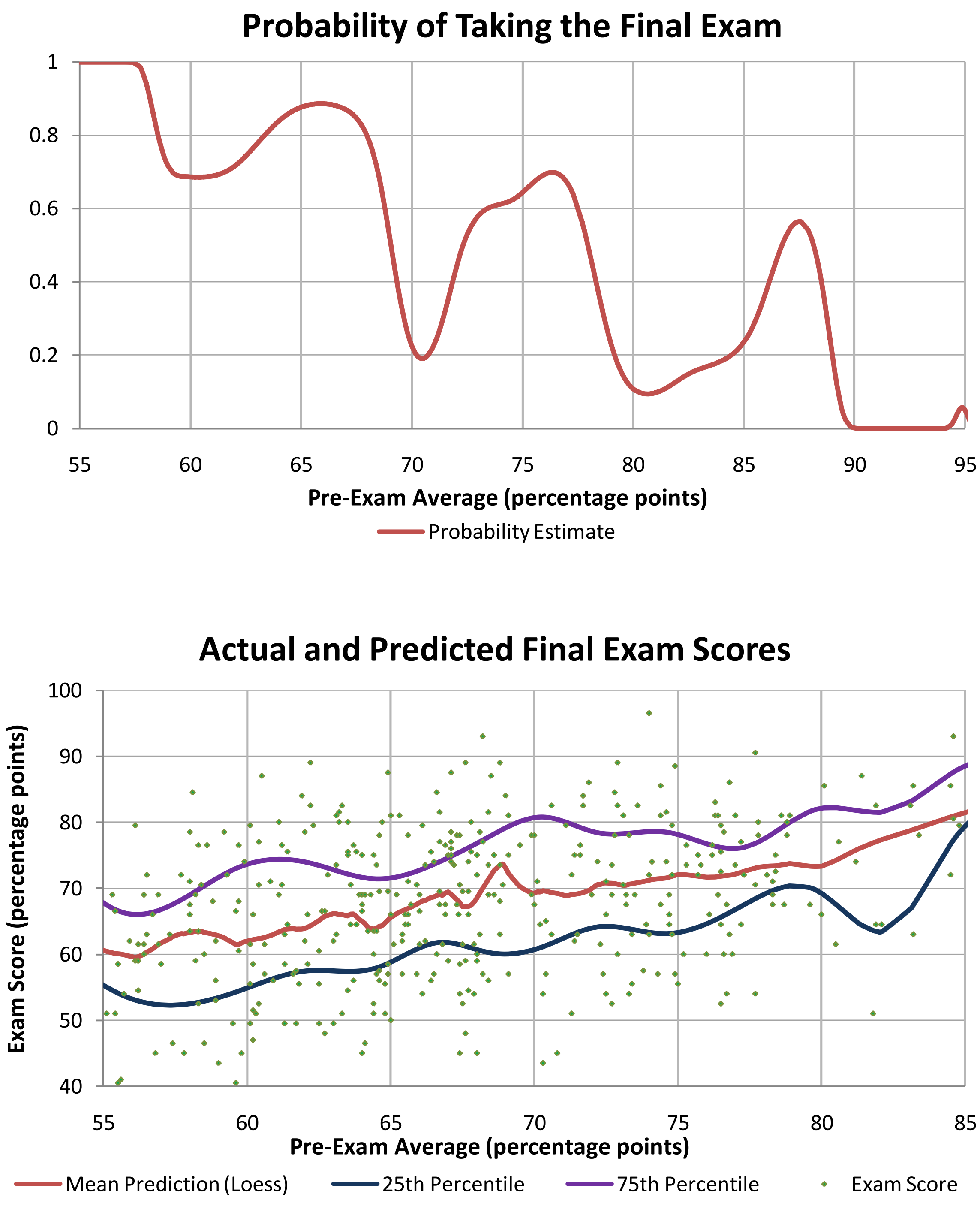

Exam Score: Deviation from Trend
Exam Score Deviation (percentage points)
4
2
0
-2
-4
55
60
65
70
75
80
85
90
95
Pre-Exam Average (percentage points)
Loess Estimate
Upper 95% Confidence Interval
Lower 95% Confidence Interval

Figure 7. Abbreviated Results Portfolio: Sweeney.



Probability of Taking the Final Exam

Actual and Predicted Final Exam Scores

Table 1. Summary of Academic Studies of Threshold Incentive Effects.

Topic	Selected Studies	Threshold	Theory	Evidence
gaming of bonus systems	Healy (1985), Courty and Marschke (2004)	annual cutoff for meeting quotas to qualify for bonuses	emphasizes potential adverse effects of thresholds	timing of reported output is adjusted to maximize bonuses
criminal behavior, drunk driving	Friedman and Sjostrom (1993), Grant (forthcoming), Iyengar (2008)	zero tolerance thresholds of various types	emphasizes potential adverse effects of thresholds or threshold reductions	reduced BAC thresholds do not effect the amount of drunk driving by youth; criminals on their “third strike” commit more severe offenses
biodiversity loss	Perrings and Pearce (1994), Muradian (2001)	where species populations are sufficiently depleted that “the ecosystem loses resilience”	emphasizes risk avoidance in a dynamic, uncertain environment	“there is abundant evidence of…threshold effects as the consequence of human perturbations on [ecosystems]”
instructional effort by schoolteachers	McEwan and Saltibanez (2005), Reback (2008)	“points” required for promotion or for passing a high-stakes test	emphasized Property 1 defined below	instructional effort appears to be stronger for those teachers or students near the threshold
analyst / publication bias in political economy, labor economics, and sociology	Tufte (2006), Gerber and Malhotra (2008)	the t values required for statistical significance of regression coefficients	formally derives the “caliper test”	researchers’ methodological choices and/or editors’ acceptance decisions favor rejections of the standard null
grade incentives on study effort	Oettinger (2002), this study	letter grade cutoffs	see the text of this study	see the text of this study

Table 2. Simulation Results.

Effort Provision. Each cell contains efficient effort, average effort under a threshold, and effort under direct measurement.

Information. Each cell contains the standard deviations of *t* among passers and nonpassers under a threshold, followed by that conditional on *T* under direct measurement.

γ = 0.175	σε = 1	σε = 2	σε = 3	σε = 4	σε = 5	σε = 1	σε = 2	σε = 3	σε = 4	σε = 5
P = 1	1.013 0.742* 0.406	**2.083** **0.131** **0.000**	3.017 0.000 0.000	3.960 0.000 0.000	4.844 0.000 0.000	1.542 2.633 0.948	2.110 2.355 1.662	2.464 2.464 2.116	2.621 2.621 2.392	2.723 2.723 2.563
P = 2	5.101 1.617 4.494	6.622 1.570 4.503	7.343 1.008 3.355	7.951 0.258 2.077	8.805 0.000 1.168	1.304 3.742 0.948	*1.338* *3.531* *1.662*	*1.823* *3.123* *2.116*	2.454 2.783 2.392	2.723 2.723 2.563
P = 4	12.949 2.256 12.342	14.008 2.832 11.890	14.665 2.976 10.676	14.555 2.737 8.681	14.487 2.262 6.851	1.148 3.849 0.948	*0.994* *3.571* *1.662*	*1.153* *3.233* *2.116*	*1.492* *3.098* *2.392*	*1.876* *3.017* *2.563*
P = 7	24.932 2.517 24.324	23.877 3.496 21.758	22.846 4.081 18.857	21.940 4.307 16.066	21.201 4.255 13.565	**1.054** **0.227** **0.948**	**0.820** **1.624** **1.662**	**0.924** **1.848** **2.116**	**1.189** **1.944** **2.392**	**1.491** **2.067** **2.563**
P = 10	27.489 2.638 26.881	28.862 3.821 26.743	27.627 4.639 23.638	26.123 5.100 20.249	24.963 5.272 17.326	**1.002** **0.211** **0.948**	**0.731** **0.504** **1.662**	**0.822** **0.829** **2.116**	**1.070** **1.167** **2.392**	**1.347** **1.493** **2.563**

Note: Bolded cells indicate an improvement in the assumed objective (efficiency in effort provision or accuracy in information provision) under a threshold system; italicized cells indicate a partial improvement—information is more accurate for passers, or for nonpassers, but not both. * means that efficiency falls despite higher average effort, which is achieved by overexertion by those near the extensive margin.

Description of Simulations: Ability is assumed to be distributed normally with mean zero and a standard deviation of three units. The threshold is placed at zero. (This need not be the optimal location, which would depend on the relative weights placed on effort provision and information accuracy. This question can be deferred for the primary purpose of these simulations--to show that thresholds can have desirable normative properties.) All other parameters are as listed in the table. Effort, actual performance, and the distributions of true and measured performance are calculated for 0.1 unit intervals of ability, for all ability levels between -10 and 10, using the results in footnote 2 and numerical computation of the extensive margin under a threshold. From this the values presented in the table are computed. While γ is fixed in these simulations, many others not reported here demonstrate that its effect on effort across the full ability distribution is quite similar to that of an appropriate change in **P**. Thus varying both parameters would be somewhat superfluous, and far more complex to present.

Table 3. Course Characteristics and Sample Sizes.

	Instructor				
	Berg	Grant	Green	Hegwood	Sweeney
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
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% to 25%	25% to 40%	25%	15% to 25%	0% to 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

Note: Prof. Green allows students to drop any test except the final exam. In the empirical work, the pre-exam average for Prof. Green's students accounts for this dropped test.

Table 4. Summary of Formal Hypothesis Tests (p-values).

	Caliper Test: Final Averages	Caliper Test: Transition Rates	Nonparam. Regression: Exam Score, Deviation from Trend	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)**	the full nonparametric estimate of the effect of the pre-exam average is not different from zero (using the best fitting smoothing value) **(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:	0.189	0.229	0.739	0.821
Berg	0.778	0.688		0.509
Grant	0.673	0.222	0.751	0.499
	0.939	0.845		0.263
Green	0.435	0.054	0.119	0.723
	0.997	0.784		0.925
Hegwood	0.795	0.983	0.276	0.339
	0.569	0.046		0.986
Royal	parametric test:	----	----	0.122
(from Oettinger, 2002)	p = 0.01 at most			0.047