

THE SIMPLE ECONOMICS OF THRESHOLDS: GRADES AS INCENTIVES¹

Darren Grant

Department of Economics and International Business
Sam Houston State University
Huntsville, TX 77341-2118

William Green

Department of Economics and International Business
Sam Houston State University
Huntsville, TX 77341-2118

Abstract: Education is one of a small but important class of goods that are created through the interaction of producer (teacher) and consumer (student), requiring complementary incentives to both parties. This paper examines how grade incentives affect student performance across a variety of courses at two regional universities, using the discrete rewards offered by the standard A-F letter grade system on final exam effort. Economic theory makes five predictions about the link between reward and effort, only one of which is trivial. Surprisingly, all are rejected in our data. These grade incentives do not influence student effort appreciably.

JEL Codes: I21, A22, D10

Keywords: educational assessment, thresholds, behavioral incentives

¹ We are extremely grateful to the three instructors who generously shared their gradebooks with us: Doug Berg, Natalie Hegwood, and Linda Sweeney.

Developing knowledge through schooling requires the joint efforts of the teacher and the student, the producer of educational services and 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 the graduates of each institution, necessitates that this authority be used to generate more effort than the student would privately prefer to give.

In schools this authority is invested partially, and in college almost wholly, in grades, which determine whether the student receives credit for the course and signals her level of performance if successful. Yet despite their importance, the availability of data, and their many interesting economic features, the incentive properties of grades 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.

Theoretically, the typical American grading system has two curious features. The first is that grades are awarded in discrete units of A, B, C, D, and F, making the marginal benefit of improved performance highly nonmonotonic. The second is that course credit is binary: full credit is awarded for a grade of D or higher, and none for a grade of F, though an A student has presumably learned much more material than the D student did, and could be awarded more credit. Both of these features are examples of thresholds, a common feature of economic life, whose properties have not been fully developed. We introduce a simple model of thresholds that generates a sequence of large set of robust predictions that can be applied to a wide range of economic activity, including the study effort induced by grades, and tested explicitly or heuristically.

Empirically, these thresholds provide leverage with which to infer the effects of grade incentives on student effort. Accordingly, we use a variety of parametric and nonparametric methods

to test the hypotheses generated by theory and infer the effect of grades on effort at all four thresholds (A/B, B/C, C/D, D/F) across the full distribution of student motivation for several college courses at two universities. The results consistently indicate that the A-F grading system has poor incentive and informational properties. In these data, grades do not motivate students to study.

I. The Behavioral Effects of Thresholds.

Public and private entities frequently must measure, and hence reward, behavioral outcomes, or “performance.” While initial measurements of these outcomes are commonly made on a continuous scale, often the consequences—the information that is released to the market, or the administratively pre-determined rewards—are binary, linked solely to passing a threshold. When viewed in marginal terms, this feature of measurement and reward represents a major departure from standard economic assumptions. The marginal benefit of improved performance is nil until one is about to cross the threshold, after which it is nil again. When the initial measurement is imprecise, so that the passing of the threshold is uncertain (conditional on performance), expected marginal benefits still rise rapidly and fall again in the neighborhood of the threshold. Both contrast with the constant or smoothly declining marginal benefits typically assumed in economic theory.

Yet thresholds are everywhere, even when a continuous system of measurement and reward appears feasible. Zero tolerance laws of various types establish a threshold for criminality above which a fixed penalty is applied (de facto or de jure) regardless of the extent of illegality. (The well-recognized perverse effects of these laws have engendered the expression “may as well be hanged for a sheep as a lamb.”) In the labor market, occupational licensure requires passing a test, whose scores

remain hidden from the public. In finance, bonds' ratings are identified on a series of thresholds, the most important of which separates junk from investment grade. And in education thresholds are everywhere—used to assess students, for course credit; teachers, for tenure; and schools, for conformity with No Child Left Behind. Despite this, the complete normative and positive properties of thresholds have not been established. The theory that follows in this section and the next identifies possible justifications for using thresholds and derives five predictions about their behavioral effects.

Behavioral Effects in Theory. Let there be a behavioral outcome of interest, t , that is additive in endowed “natural ability,” v , and effort, f : $t = v + f$, which is measured precisely and valued by the market at price \mathbf{p} . Values of v are distributed normally throughout the population with standard deviation σ_v . Each individual's effort is chosen to maximize the difference between the rewards from effort and its cost: $\mathbf{p}f - C(f)$. The solution, $f^* = C'^{-1}(\mathbf{p})$, is also the efficient outcome as long as the price \mathbf{p} is appropriate (there are no externalities, for example). Continuous measurement provides ideal information to users and appropriate effort incentives: thresholds are not needed.

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 typical 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, in proportion to the variance of each. Let $T = t + \epsilon$, where ϵ is error is

measuring the true outcome, independently and normally distributed. The market price of a unit increase in T generated by effort is $\mathbf{p}\sigma_v^2/(\sigma_v^2+\sigma_\epsilon^2) < \mathbf{p}$, so each individual underprovides effort.² 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 $T \geq 0$, that is, $I(T \geq 0)$. The market value of passing the threshold equals $\mathbf{P} = (\bar{\tau}_{\text{passers}} - \bar{\tau}_{\text{nonpassers}})\mathbf{p}$; the probability of passing the threshold, conditional on effort, is now $\Phi(v+f)$, where Φ is the cumulative distribution function for ϵ . The expected marginal returns to effort are now bell-shaped, centered around $t = 0$. 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 ; the vertical axis marginal costs and expected marginal benefits, in arbitrary logged units. (Taking logarithms simplifies the graphics and corresponds to the parametric version of the model put forth shortly.) Five students, A-E, indexed by i , are represented on the graph, their upward sloping marginal cost of effort lines beginning at v_i . For sufficiently low v , as for student A, there is no point of intersection between marginal costs and marginal benefits, 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, as in the figure, be reached where $t < 0$; if so effort increases until

² This is a symmetric sub-game perfect Nash equilibrium to an N-person “effort game,” where each person’s effort is optimal given the choices of everyone else. Because each person provides the same amount of effort, the variance of t ex post equals the variance of v ex ante.

the intersection reaches the vertex of the parabola, as for student C, and declines steadily thereafter (student D) until, at sufficiently high, positive v , it returns to nil (student E). Those for whom $0 < v_i < v_E$ will probably pass even without effort, but assessment is uncertain so they put forth “precautionary” effort to increase their chances.

The resulting (v,f) and (v,t) loci, in Figure 2, reveal five positive predictions of the theory.

1. *Effort is non-monotonic: least for individuals far from the threshold ($v \ll 0$), greater for those close to it ($v \approx 0$), and greatest for those in between.* This is a consequence of the extensive margin, which is itself a consequence of the non-monotonic returns to effort.
2. *Effort rises more quickly than it falls; that is, in Figure 2a, the (v,f) locus takes a sawtooth shape, rising faster on the left (line BC) than it falls on the right (line CE).* Along with the extensive margin, at which effort increases discretely, this follows from an upward sloping marginal cost of effort intersecting with a symmetric parabola.
3. *Effort is positive at $v=0$.* Error in assessing t motivates precautionary effort to increase the individual’s chances of passing.
4. *Effort does not fully compensate for endowment, that is, $\Delta f/\Delta v > -1$.* Individuals with better endowments work less hard and still have better outcomes.
5. *At the endowment associated with maximum effort, the probability of passing the threshold equals or exceeds $1/2$; that is, line OC in Figure 2a has a slope ≤ -1 .* An interior maximum, as in Figure 1, is reached where $t=0$, so $f = -v$ and line OC has a slope of -1 . Otherwise maximum effort occurs at the extensive margin, where by property 4 above $f \geq -v$, and line OC is more steeply (negatively) sloped.

None of these properties depend on functional form, only an upward sloping marginal cost of effort and a symmetric, mean zero error distribution, both of which are unexceptional.

Behavioral Effects in Practice. Do individuals respond to thresholds? We have eclectic evidence that they can. Three examples are provided in Figure 3, using three different formats.

The top graph in the figure illustrates the response to the October 17, 2005 change in the

bankruptcy law, which made filing for bankruptcy more difficult and less attractive. This is a temporal threshold, on the filing date; thus this graph adopts the $\{v, t\}$ format of Figure 2b, with the directionality of v reversed (earlier is better). Because one measures whether the threshold is met with precision, the prediction is for filings, which were roughly uniform in the preceding years, to increase steadily until the law changed and then plummet, which clearly is what happens.

The middle graph, from Reback (2008), shows the effect of thresholds on effort provision by Texas schools and teachers. This graph adopts the $\{v, f\}$ format of Figure 2a. In Texas, schools' ratings were based on the fraction of students reaching a fixed passing standard on the TAAS (Texas Assessment of Academic Skills) test, which was given to almost every grade. One can thus forecast the likelihood that a student in a given year is likely to pass the TAAS, from prior years' performances, and to focus educational resources on those previously-failing students that are most likely to pass this year. Reback's graph clearly shows properties 1-4 above (property five cannot be tested, because the two axes are in different units).

The bottom graph, from Tufte (2006, p. 144), illustrates the effect of the $p=0.05$ threshold in statistical hypothesis testing in economic research, in yet a third way—with a distribution of observed outcomes T . In Tufte's words:

Consider a research issue between those who believe there is no relationship between two variables [H_0] and those who believe there is a relationship (H_A)... Both schools of advocates seek to publish decisive results; and both seek to avoid ambiguous results. Such advocacy may lead to evidence selection, resulting in...a heaping of results that strongly favor either H_0 or H_A , and fewer results [in between]. I compiled the distribution of the 248 published t-statistics from all 17 of the then-published studies of election year macroeconomic conditions and the U.S. national vote for the political party of the incumbent president. Both the H_0 and H_A heapings, as well as the zone [in between], are [visible].

Our empirical work presents results in all three of these formats, plus one other, to make our case.

II. Adopting Thresholds.

We are now prepared to examine three reasons a threshold might be actively preferred to a system of direct measurement. To do this it is valuable to specify costs as $C(f) = k \cdot \exp(-\gamma f)$. Retaining the normal distribution of ϵ , this model solves analytically for effort and performance, so the efficiency and informational properties of thresholds in a population can be easily determined (numerically). Effort, conditional on ability v , depends on three parameters: $\{P/k\}$, the value of passing the threshold, now normalized relative to costs; σ_ϵ , imprecision in measuring t ; and γ , diminishing returns or “fatigue” in the provision of effort. Simulation results, used to illustrate the points made below, are presented in Table 1 for a reasonably broad set of parameters, retaining the zero threshold and central, normal distribution of v (with a standard deviation of three units).

Signaling. Spence (1973) showed that an educational threshold can be useful, aside from any human capital development, by providing valuable information to employers about workers’ underlying aptitudes (v in our model). What Spence did not show was that a threshold is an optimal way to do this. This is because it is not: direct measurement, even if imperfect, is always superior, because it does not throw away valuable information on which to condition, as does the threshold.

Motivating. Previously we showed that effort is underprovided under direct but imprecise measurement of outcomes: the expected marginal benefit of additional effort is attenuated, because some of the additional effort is inferred to be noise, instead, in the solution to the signal extraction problem. Furthermore, this offset is constant: actual effort equals efficient effort minus $\sigma_\epsilon^2/\sigma_v^2$, so that the reduction in effort can be large in relative terms when the efficient effort level is modest.

Under these circumstances, thresholds can increase effort and improve efficiency, by

intensifying the effort of individuals near the threshold. Given $\mathbf{P} = (\bar{t}_{\text{passers}} - \bar{t}_{\text{nonpassers}})\mathbf{p}$, the expected rewards to passing the threshold can be quite large. The simulations in Table 1, however, show that this is difficult to execute. Under direct measurement all individuals provide some effort; under the threshold, those who are below or far above the extensive margin provide little or no effort, while others, just above the extensive margin, may overprovide effort. As a practical matter, thresholds appear to increase efficiency in effort only when agents are sufficiently unmotivated under direct measurement that they provide virtually no effort at all. This is a Pyrrhic victory.

Informing. This is distinct from signaling because the trait of interest, the outcome t , is under the agent's control. Those interested in knowing an agent's performance, such as employers, may value more accurate information about t —which, surprisingly, can be provided by using thresholds.

This is because, with properly placed thresholds, effort is negatively correlated with ability: $\text{cov}(v, f) < 0$. Many individuals with negative v put forth considerable effort to pass the threshold, while those with positive v put forth a little precautionary effort or none at all. This is not so with direct measurement, where (in our scenario) $\text{cov}(v, f) = 0$. The existence of a threshold, and with it the extensive margin, bifurcates individuals into two groups, with disparate average outcomes *across* groups, and similar outcomes *within* groups. When t is measured imperfectly, there can be significantly less error variance here than with direct measurement: that is, $\text{var}(t|T \geq 0) < \text{var}(t|T)$ and $\text{var}(t|T < 0) < \text{var}(t|T)$.

This is confirmed with the Table 1 simulations, which also show that this system works best when the market rewards to passing the threshold are largest (higher \mathbf{P}) and when fatigue is smallest (lower γ), which both yield more effort. As before, when discussing motivation, error variances initially decrease in σ before rising again: imperfect measurement of T may be actively beneficial.

A potential theoretical explanation for thresholds in grading can now be offered, which combines imprecision in the measurement of outcomes with the effect of incentives on the production of human capital by students: when these incentives are weak, a threshold grading system can augment effort and thus improve efficiency; when these incentives are strong, such as system can improve the performance information provided to employers or other educational institutions.

III. Incentives and Thresholds in Education.

We now focus on education specifically, to see what the historical record and existing research indicate about thresholds, their behavioral effects, and their normative properties.

Historical Evidence on Threshold Adoption. The previous section identified two potential reasons for adopting a threshold grading system. We can turn to the historical record to see which of these motivations resulted in the development of the letter grade system so widely used today. The answer is: neither of them.

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. They were extended during the Industrial Revolution, as the state tried to exercise more control over universities' examination processes, including entrance examinations, to improve the quality of its civil servants, which 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 a student's ability at rote memorization. Rudolph (1977, p. 145) points out these were mostly just "gestures in public relations," as the examinations were not rigorous. The first full system to rank students appeared at Yale in 1785, using four tiers, as in English universities. This was modified to a four-point numerical scale in 1813 that included both whole numbers and decimals. Harvard adopted a similar approach at roughly the same time, expanding its twenty-point grading scale to one hundred points in order to measure achievement more exactly.

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. As Harvard's new curriculum and teaching methods spread throughout American higher education, so did its new grading system. We do not know why the numerical system was changed to thresholds, however.

Similar changes in grading occurred in public schools, however, for better-documented reasons. Both enrollment and professionalism in the public schools increased dramatically during the late 19th century, and with it came a shift in assessment from written narratives to parents to percentages on examinations in different subject areas. Then, in 1912 and 1913, Wisconsin researchers Daniel Starch and Edward Charles Eliot (1912, 1913) seriously challenged the reliability of percentages as indicators of achievement. They found 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. This led to a gradual movement away from percentage scores to scales with fewer, larger categories, such as the "Excellent," "Good," "Average," "Poor," and

“Failing” system that grew into today’s A-F scale.

In summary, grading systems evolved with the educational system and 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, and second, motivated by Starch and Eliot, to mask the disparity in instructors’ grading standards. Any beneficial properties of threshold grading systems would be purely incidental.

Behavioral Effects of Grade Incentives. This history raises suspicion about the effectiveness of grade incentives in education. So too does current evidence that complements the historical record.³

The educational psychology literature initially emphasized a model in which behavior responded to external reinforcements, such as grades, and in which these reinforcements could be adjusted in almost Keynesian way to bring about the desired behavior. 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. The net result is a view of incentives that, while not at odds with the traditional, economic, instrumental view, casts a much wider net.

The most important conclusion of this research is that extrinsic motivation and intrinsic motivation are substitutes: strengthening one weakens the other. This diminishes the net effect of

³ This discussion relies on two recent assessments of the field, Stipek (1996) and Elliot and Ista (2008). Grade incentives receive even less attention in the literature on educational assessment. The journal *Studies in Educational Evaluation* has thirty-four volumes. An online search of titles, abstracts, and keywords in all articles for the word “incentive” yields a single match, which is not relevant to the topic of this paper.

extrinsic rewards on student achievement. In particular, students often have an intrinsic “achievement motive” that is weakened by the use of incentives. Furthermore, extrinsic incentives’ effects are influenced by students’ perceptions of competence and self-efficacy. If these are poor, students adopt a “performance-avoidance” goal, which is essentially a maximin objective function that attempts to moderate bad outcomes rather than strive for good outcomes. 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 unresponsiveness of college graduation rates to the rapid rise in the college wage premium. As it stands, though, the economics literature contains very few studies exploring the effects of grade incentives on student achievement, focusing more on the determinants and ramifications of variation in grading standards across teachers, fields of study, or institutions. There is some evidence more difficult instructors impart more knowledge (see Grant, 2007, and sources within), though this might have more to do with teaching methods than incentives, which are not distinguished in these studies. The second-most pertinent study is Farkas and Hotchkiss (1989), who measure the degree to which grades respond to student effort in various classes, and use these differences to explain effort differentials across student groups. Their main finding is that incentives have small effects on behavior.

The most pertinent study is Oettinger (2002), who explores the effect of grade thresholds on final exam performance for college students, as we do, and concludes that they matter. This paper and his differ in three main respects: 1) we emphasize economic significance, and Oettinger statistical significance; 2) Oettinger’s estimates are parametric, and ours mostly nonparametric; and most importantly, 3) our theoretical development, unlike his, emphasizes the role of uncertainty in passing

the threshold, which leads to qualitatively different predictions. Below we compare Oettinger's estimates to ours, and argue there is less dissonance between them than appearances suggest.

IV. Data.

The data used in this analysis was generously provided by five university instructors teaching four different courses, both upper and lower division, at two Texas universities during a subset 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 contained in Table 2.

University grading systems are typically either norm-referenced (relative grading) or criterion-referenced (absolute grading). Students in norm-referenced systems are evaluated relative to one another; while thresholds still separate letter grades, these thresholds are not specified in advance, so students are unlikely to be motivated by them. Criterion-referenced grading, in contrast, 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 course, the data contain all test scores and homework grades recorded in the course, and we know 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. Typically 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 approached 1,050 over the sample period; six-year graduation rates, slightly under 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.

Each instructor in our data is terminally qualified, with roughly a century of combined full-time teaching experience in 2008; in their first year in our sample each has at least four years prior experience teaching the course included in our data. Course evaluations and administrators' judgements suggest that these instructors are typically 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, without typically informing students in advance that they do this. In our data, each instructor teaches at least 655 students, so that both parametric and nonparametric estimates of effort provision, as reflected in final exam scores, can be obtained with reasonable precision.

V. Results.

Results for four instructors are presented in Figures 4-7: Professors Berg, Grant, Green, and Hegwood. Each figure contains a portfolio of results for each instructor, which illustrate the distribution of final averages, the change in these averages after taking the final exam, and final exam performance conditional on the pre-exam average. We discuss these four instructors' results collectively.

Distribution and Transition. The first graph in each portfolio is a simple frequency distribution of course averages, before taking the final exam and after. Averages, in percent, 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 possible exceptions that could result randomly: B's for Profs. Berg and Grant and D's for Prof. Green. Many students' averages change after taking the final exam, 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 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 transition matrix, with row and

column totals, and associated proportions, along the outside.

Each matrix provides three pieces of evidence about strategic final exam behavior. 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 second piece of 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.

The third piece of evidence concerns transitions for students in the middle of their grade range, in the second row of the matrix. If they behave strategically, transitions to the lower two points of their grade range, in the left bolded cell, should be more likely than those 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. Students are not motivated by grades.

Exam Scores. The next two figures in each portfolio look directly at the exam scores themselves, as a function of each student's pre-exam average. The first of these presents actual and expected exam scores themselves, while the second—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 top, multi-layered figure begins with a scatterplot of actual exam scores against the pre-exam average. The incredible (conditional) variation in individual exam scores, combining variation in effort across students, exam difficulty across semesters, luck, and extraneous factors such as exam-day health, is immediately apparent (and supports our incorporating imprecision in the theoretical model). To this scatterplot are added three smoothed sets of predicted exam scores: the mean, in the center line, calculated using a loess smoother, and the 25th and 75th percentiles, in the other two lines, calculated by applying quantile regression to the exam scores that were adjusted for inter-semester differences in exam difficulty.⁴

The pre-exam average reflects, mostly, student ability, but it too is affected by random factors, so there should be regression to the mean, which should make the slope of each line less than one. It always is. This regression to the mean 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 scatterplot, and a slight convex shape is observed for Profs. Grant and Hegwood. The long arc of the relation between dependent and independent variables, therefore, is adequately modeled with a simple quadratic. This is, in fact, actively preferred, as the the “polynomial wiggle” that could be

⁴ These differences were estimated by using the coefficient estimates on the semester dummies that were included with the loess smoother. 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. Further details are available from the first author.

introduced by a higher-order polynomial could obscure any strategic variation in exam scores.

Strategic behavior should be apparent in deviations in from this long arc, located slightly under the ten-point thresholds for each letter grade. This is apparent in 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, which presumably reflects the most grade-motivated students at each pre-exam average. And even where it is apparent, the effect is not large, nor, as it turns out, significant.

This is demonstrated in the last figure of each portfolio, which depicts (again using the loess smoother) the deviation of the 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—shortly before the grade threshold. Statistical tests of the null hypothesis that there are no deviations from trend whatsoever are provided in our statistical software, and these never reject the null, with p-values that typically exceed 0.5.

Exam Taking. The final instructor for which we have data, Prof. Sweeney, 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 Figure 8. The top graph illustrates the probability of taking the final exam, estimated nonparametrically using transformation regression, as a function of the pre-exam course average (and a dummy for whether a previous test was missed). This graph, in contrast to those preceding it, exhibits dramatic variation. It indicates clearly the diminishing marginal value

of each successively higher grade, so that 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: probabilities increase within each grade range as one approaches the grade threshold (which is shifted left by about three percentage points, because this instructor rounds up generously). This should occur when each student knows with certainty, as here, her course grade were she not to take the final. If she does take the final, prediction 4 implies her post-effort passing probability increases the closer is her pre-exam average to the threshold.

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 threshold. Exam-taking appears to respond to grade incentives, but not exam performance. Across all five instructors, there is virtually no evidence that students strategically raise their exam scores via increased study effort when their grades are most likely to benefit, even if it means the difference between passing and failing.

Estimation. A variety of nonparametric estimators and smoothing values could have been used to present the estimates in our results portfolios. It is impractical to illustrate the differences between them, but also unnecessary, as these can be summarized with two key points. First, the choice of estimator is inconsequential. We constructed estimates with transformation regression, least absolute difference regression (quantile regression at the median), and nonparametric spline estimators, in addition to the loess estimator presented above, all to little effect. Second, we chose smoothers that *exaggerated* the variation in the estimates. Under traditional choice criteria the smoothing parameter

would be much larger, and the resulting nonparametric estimates far smoother. Deviations from “trend” in these estimates rarely exceeded one-half of one percentage point, and were insignificant.

For completeness, we also present parametric estimates as well, in Table 3, using a specification introduced by Oettinger (2002), which includes a trinomial in v and interval dummies for the pre-exam distance (δ) 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 are intended to capture f , but in reverse–effort relative to those on the borderline. Accordingly, Oettinger expects increasingly negative coefficients across these four dummies. (Our theory, which differs from his in emphasizing uncertainty in the passing of the threshold, does not make this prediction, and also distinguishes between those above and those below the threshold.) Across all four instructors studied here, these four dummies are jointly insignificant, reinforcing our nonparametric estimates.

When Oettinger estimates his model on grades from a micro principles class at the University of Texas’s flagship Austin campus, he concludes that strategic effort is apparent, based on the coefficient estimates at the bottom of Table 3 and some evidence that students’ final averages cluster just above the grade thresholds. Even if these dummies are statistically significant (jointly in the least absolute difference regression), the implied effects, roughly one percentage point on average, are modest. Study effort does respond to grade incentives in Oettinger’s data, but, still, not appreciably.

VI. Conclusions.

The weak response to grade incentives in higher education has two key ramifications. First, most obviously, it places a significant limitation on instructors’ means of motivating students. This

limitation need not be incapacitating, as the education literature discusses multiple means of fostering student motivation, many of which do not require extrinsic rewards such as good grades. But it is disappointing. As educational credentials and institutional reputations carry great value, it is efficient to engender more study effort from students than they would privately prefer to give.

The second great ramification is that grades, at least individual course grades, will not provide good information about students' competence. This property goes in hand with the previous one. Highly grade-motivated students would tend to cluster just above their preferred grade threshold, making the course grade a good indicator of student achievement in that class. Such students would (probably) also have similar grade expectations across classes, making GPAs excellent indicators of student skill. But this does not happen. This helps us understand 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 "nature"—comprising random factors and other unmeasurable factors unrelated to student or teacher skill.

Producing human capital through education is a particularly confounding task for markets to solve effectively, as it requires sustained investment across time through the joint efforts of the student and a succession of teachers. Efficiency in this market is even more difficult to achieve, impeded by significant informational difficulties, subject to equity concerns, and requiring adjustment for positive externalities. The difficulty motivating students through course grades discovered here adds yet one more complexity to the mix.

REFERENCES

- American Bankruptcy Institute. "Personal Bankruptcy Filings by Quarter," available online at: http://www.abiworld.org/am/template.cfm?section=Bankruptcy_Statistics1, accessed Aug. 2008.
- 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).
- Grant, Darren. "Grades as Information," *Economics of Education Review*, 26, 2:201-214 (2007).
- Oettinger, Gerald. "The Effect Of Nonlinear Incentives On Performance: Evidence From "Econ 101,"" *Review of Economics and Statistics*, 84:509-517 (2002).
- 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).
- 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).
- 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. Theoretical Threshold Effects on Individual Effort Provision.

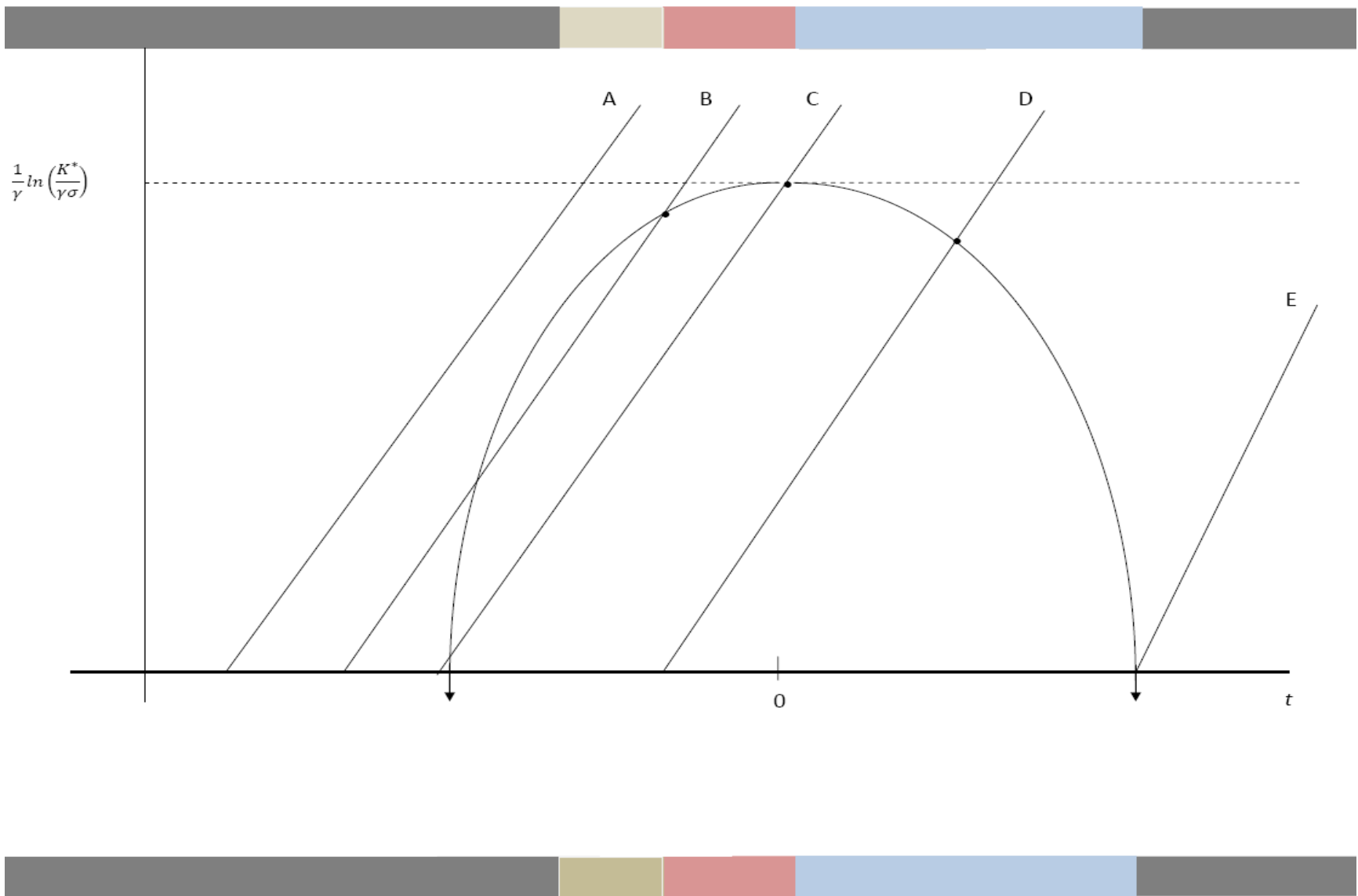


Figure 2. Theoretical Threshold Effects Across a Population with Varying Ability.

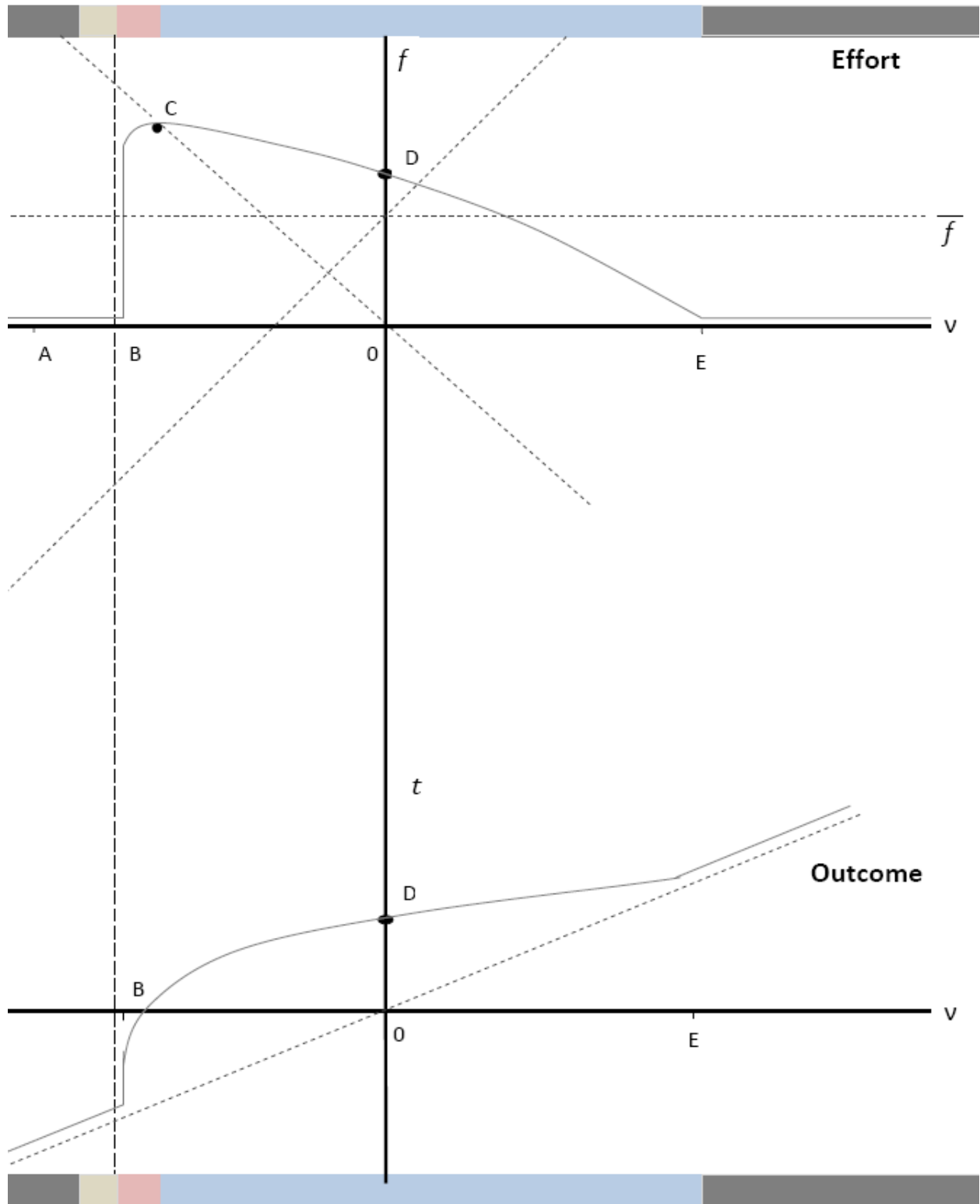
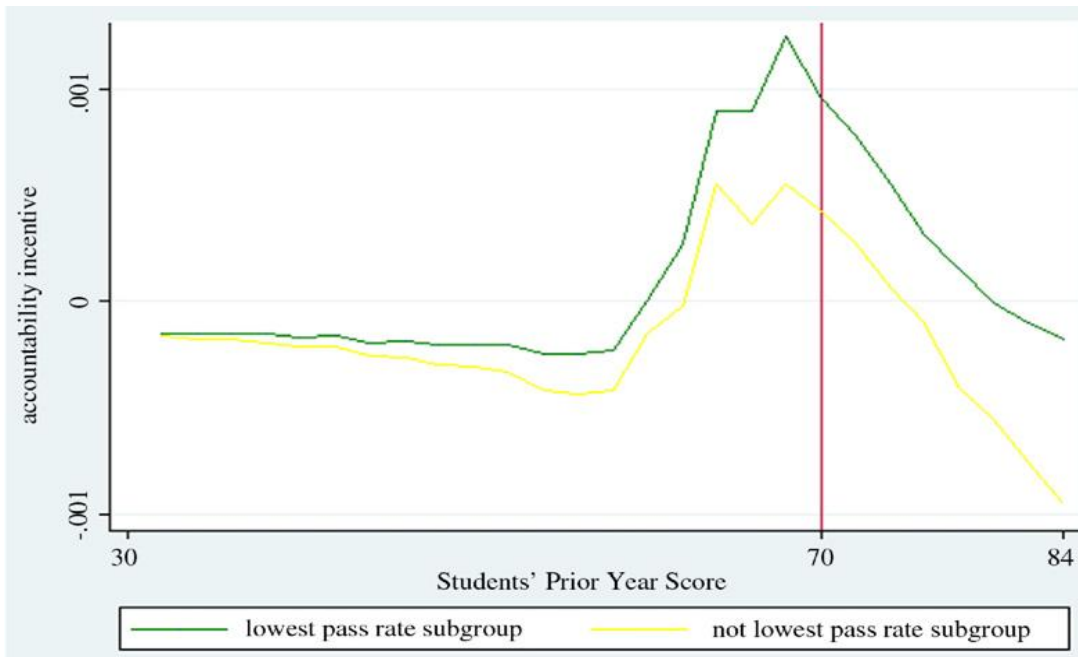
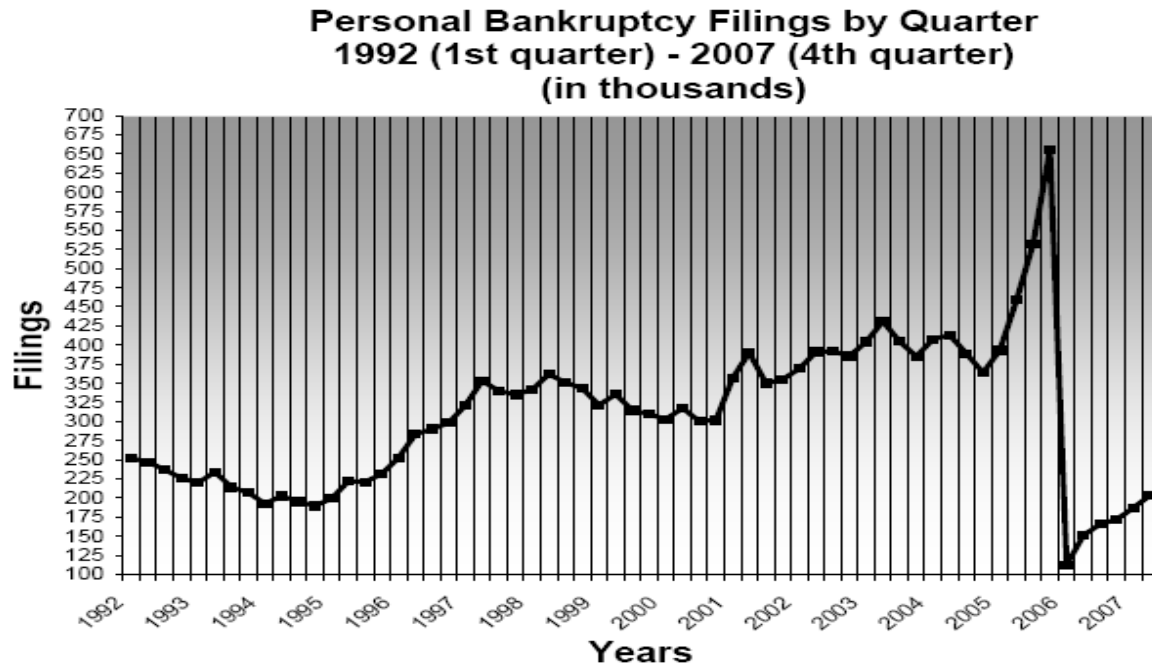
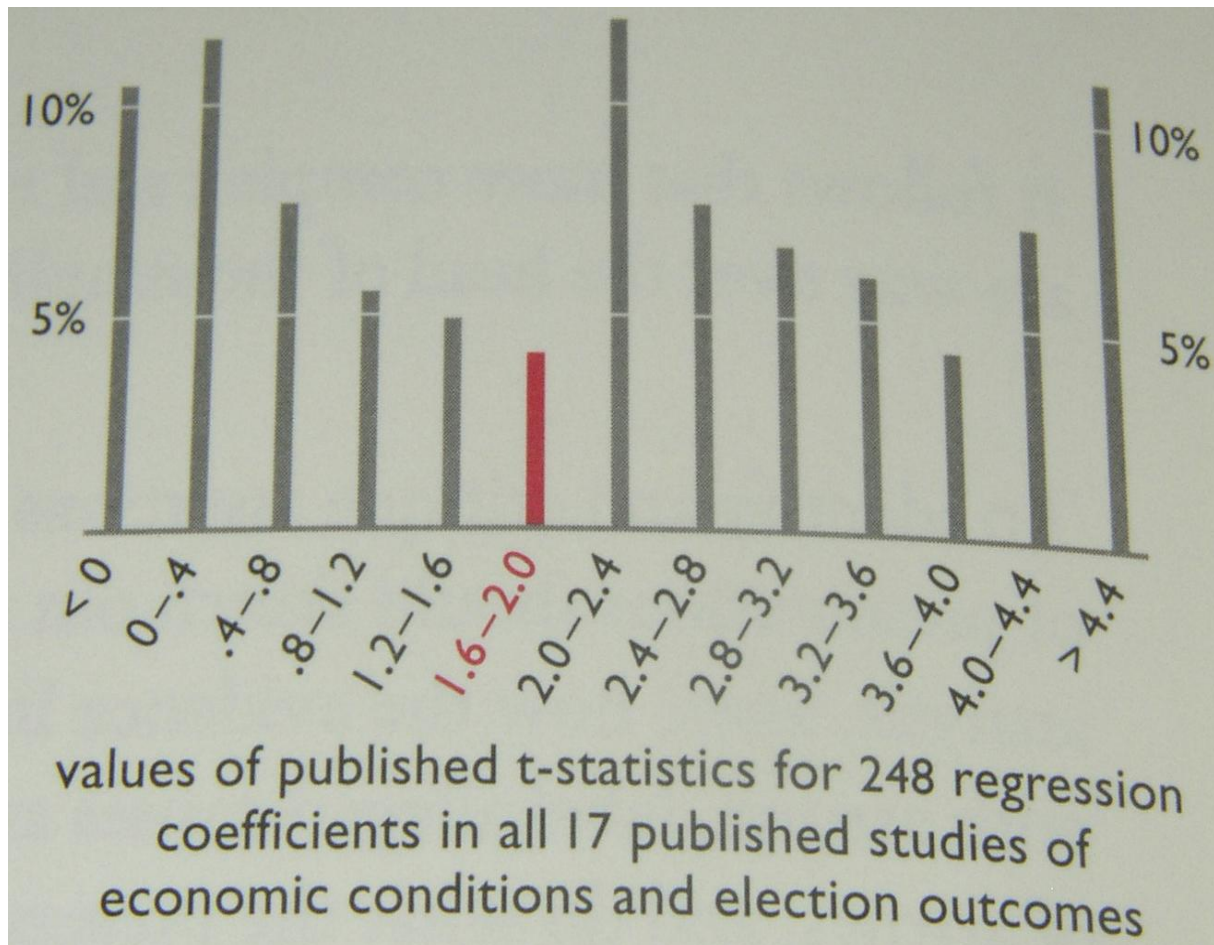


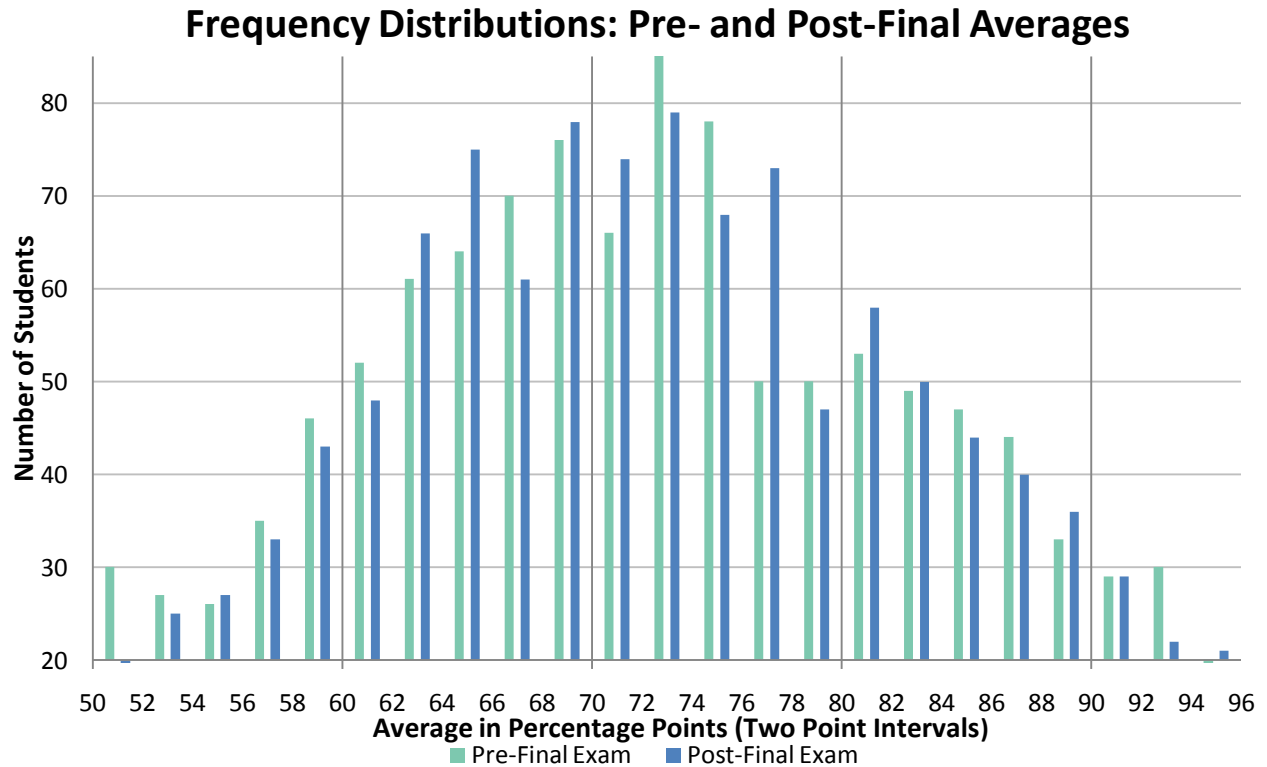
Figure 3: Three Examples of Thresholds and Their Impact on Behavior.





Note: The top graph, from the American Bankruptcy Institute (2008), depicts quarterly personal bankruptcy filings by quarter from 1992:1 through 2007:4. Bankruptcy law became more exacting on October 17, 2005. The middle graph, from Reback (2008), depicts the deviation, from expectations, in the probability of passing Texas' old TAAS test, as a function of the school's prior year's passing rate. The threshold for schools to be rated "Academically Acceptable" was seventy percent. The bottom graph, from Tufte (2006), depicts the distribution of published t-statistics on economic variables used in predicting election outcomes in all studies that were published prior to the completion of his 1978 book *Political Control of the Economy*. Standard tests of statistical significance typically require a t-statistic of 1.96 to reject the null.

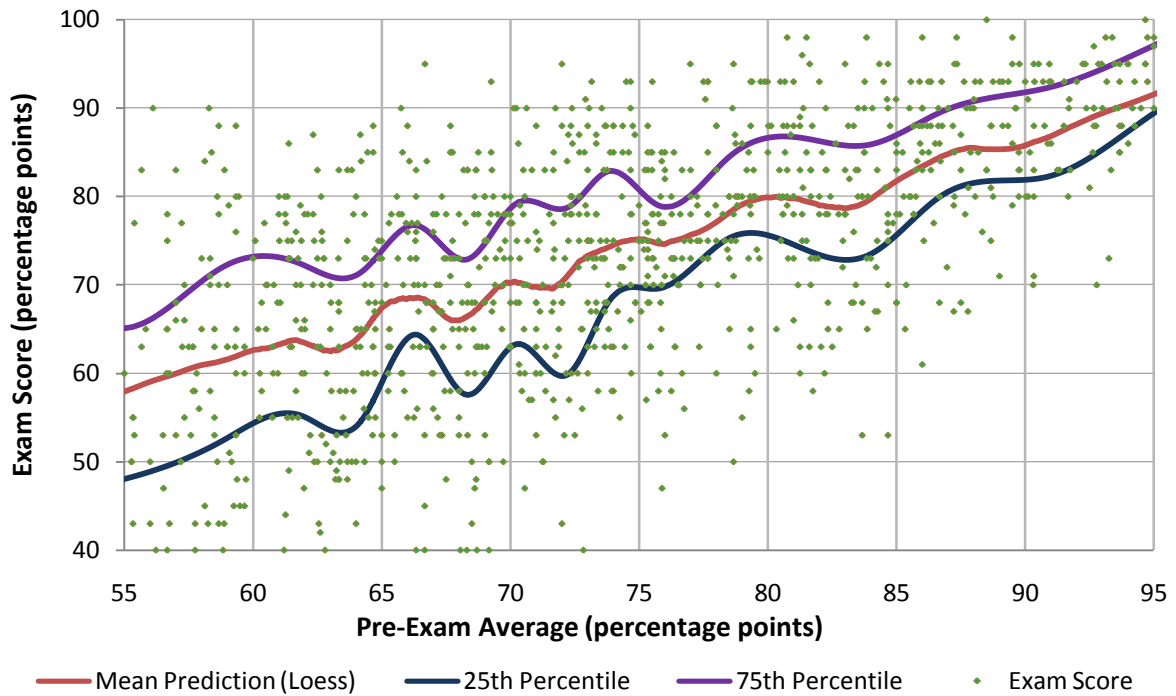
Figure 4. Results Portfolio: Berg.



TRANSITION MATRIX

Post-Final → Pre-Final ↓	Bottom Two Points of Range	Middle Six Points of Range	Upper Two Points of Range	Row Totals
Bottom 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)
<i>Column Totals</i>	<u>223</u> <u>(0.197)</u>	695 (0.614)	<u>214</u> <u>(0.189)</u>	1132 (1.000)

Actual and Predicted Final Exam Scores



Exam Score: Deviation from Trend

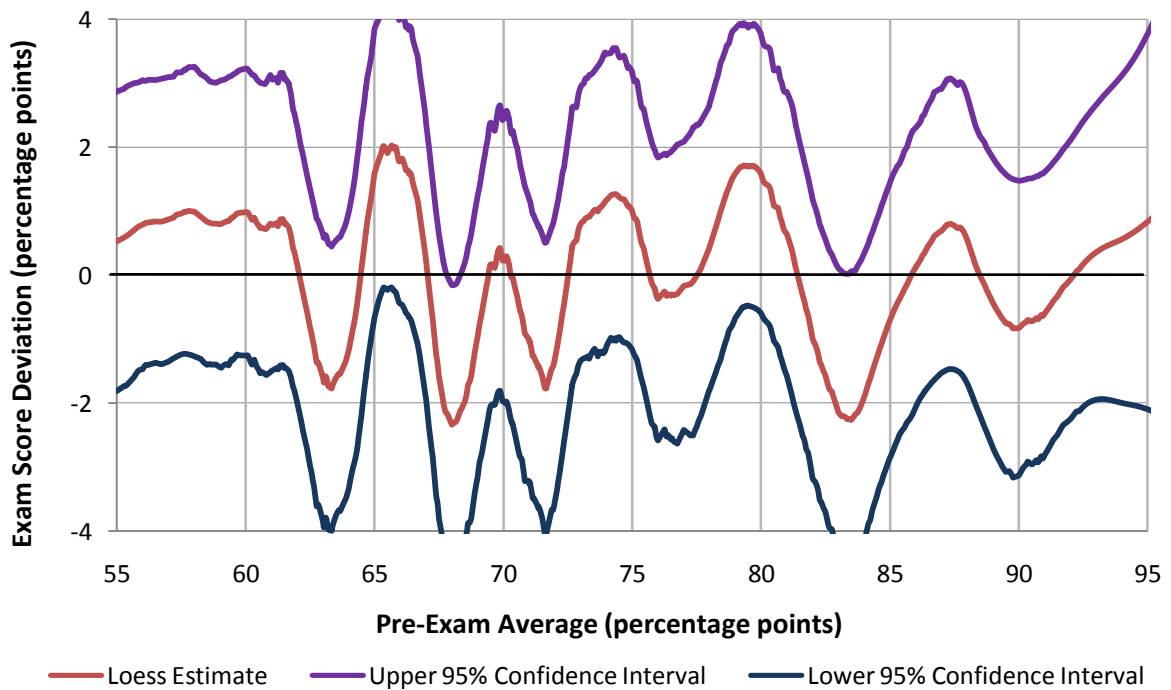
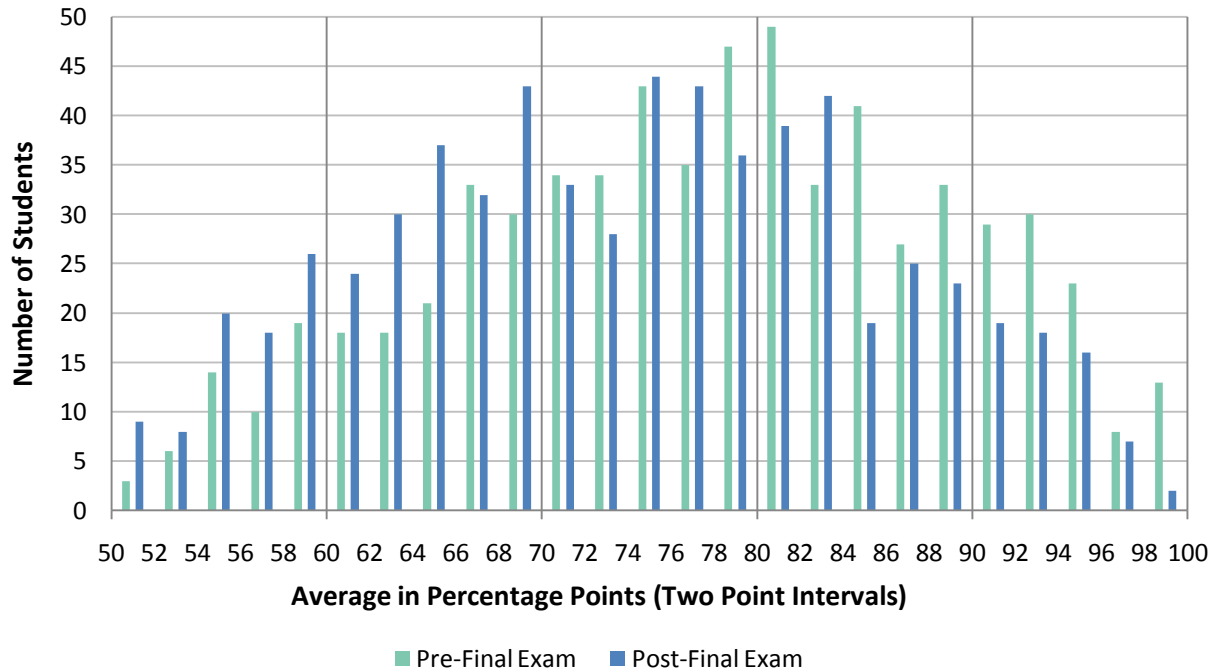


Figure 5. Results Portfolio: Grant.

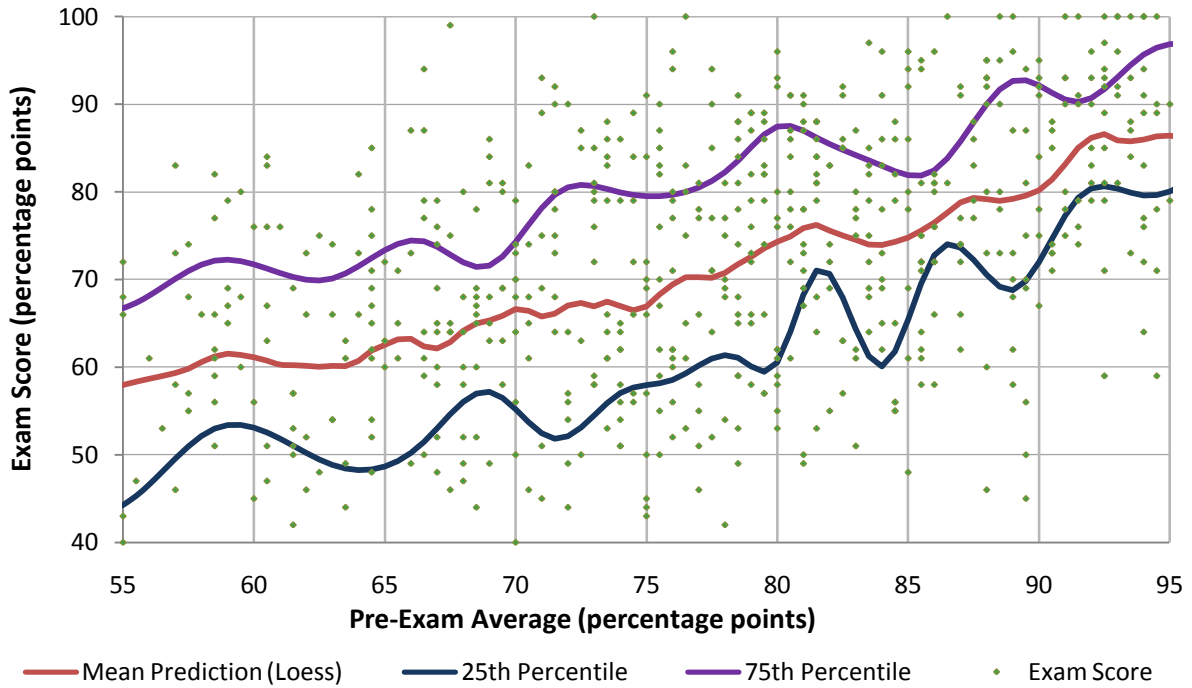
Frequency Distributions: Pre- and Post-Final Averages



TRANSITION MATRIX

Post-Final → Pre-Final ↓	Bottom Two Points of Range	Middle Six Points of Range	Upper Two Points of Range	Row Totals
Bottom 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	<u>125</u> <u>(0.19)</u>	395 (0.60)	<u>135</u> <u>(0.21)</u>	655 (1.00)

Actual and Predicted Final Exam Scores



Exam Score: Deviation from Trend

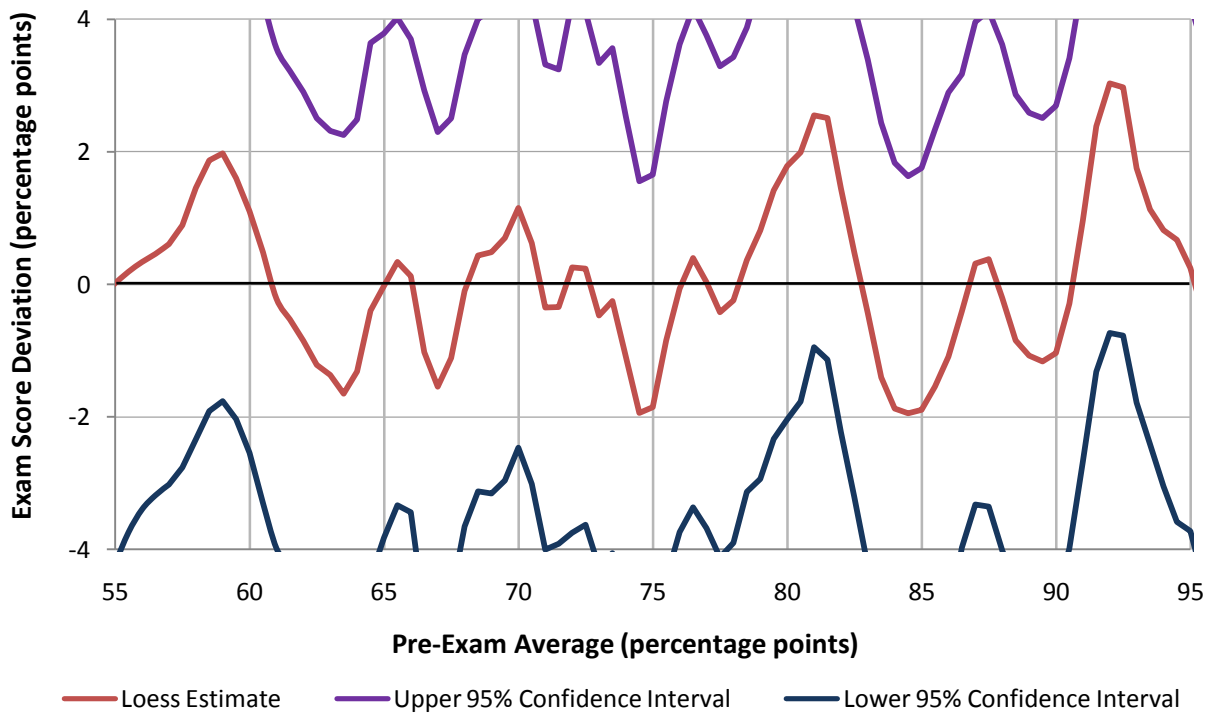
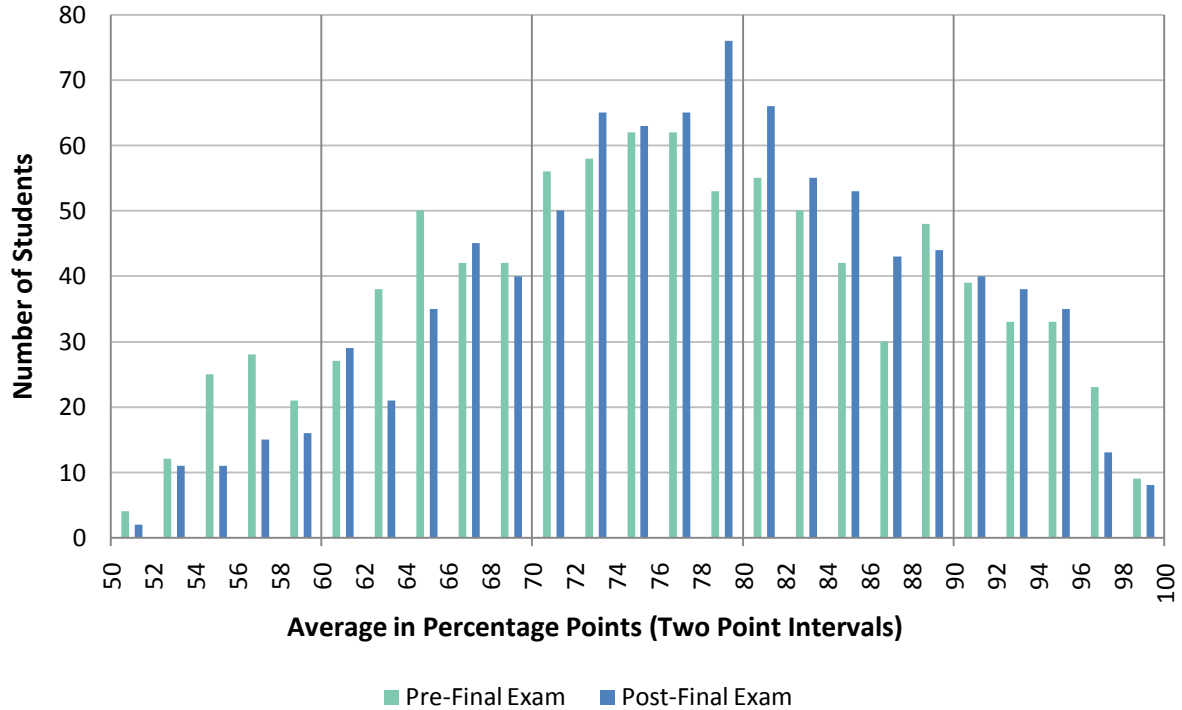


Figure 6. Results Portfolio: Green.

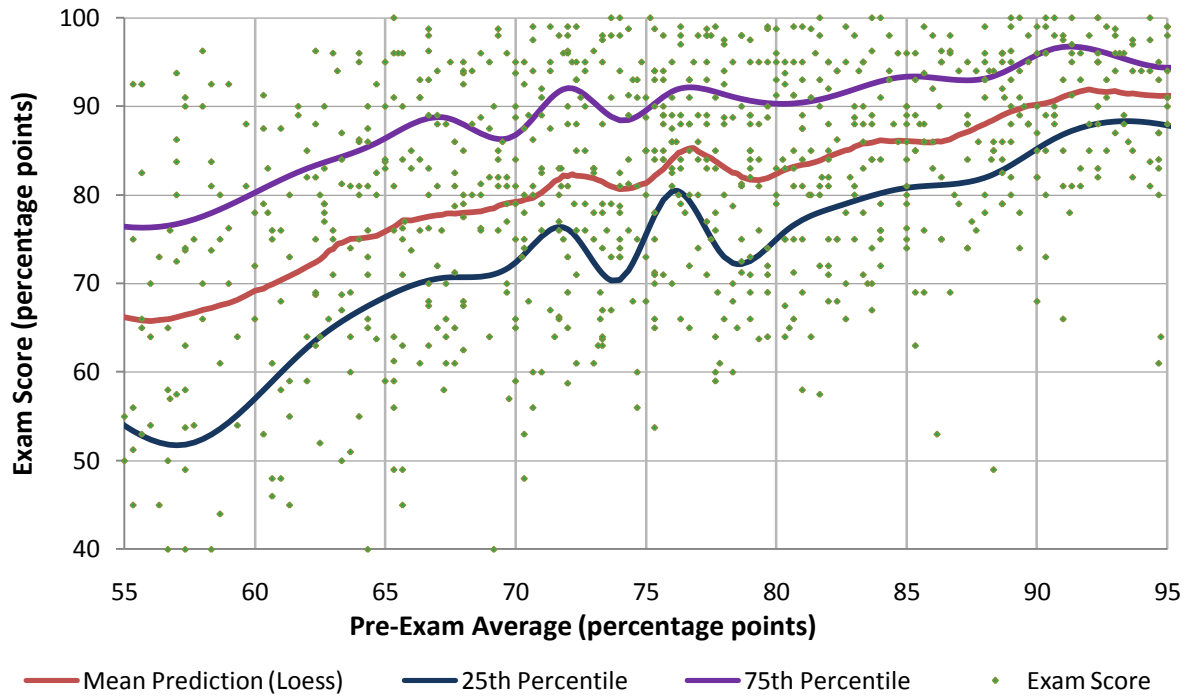
Frequency Distributions: Pre- and Post-Final Averages



TRANSITION MATRIX

Post-Final → Pre-Final ↓	Bottom Two Points of Range	Middle Six Points of Range	Upper Two Points of Range	Row Totals
Bottom 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	<u>188</u> <u>(0.20)</u>	569 (0.60)	<u>186</u> <u>(0.20)</u>	943 (1.00)

Actual and Predicted Final Exam Scores



Exam Score: Deviation from Trend

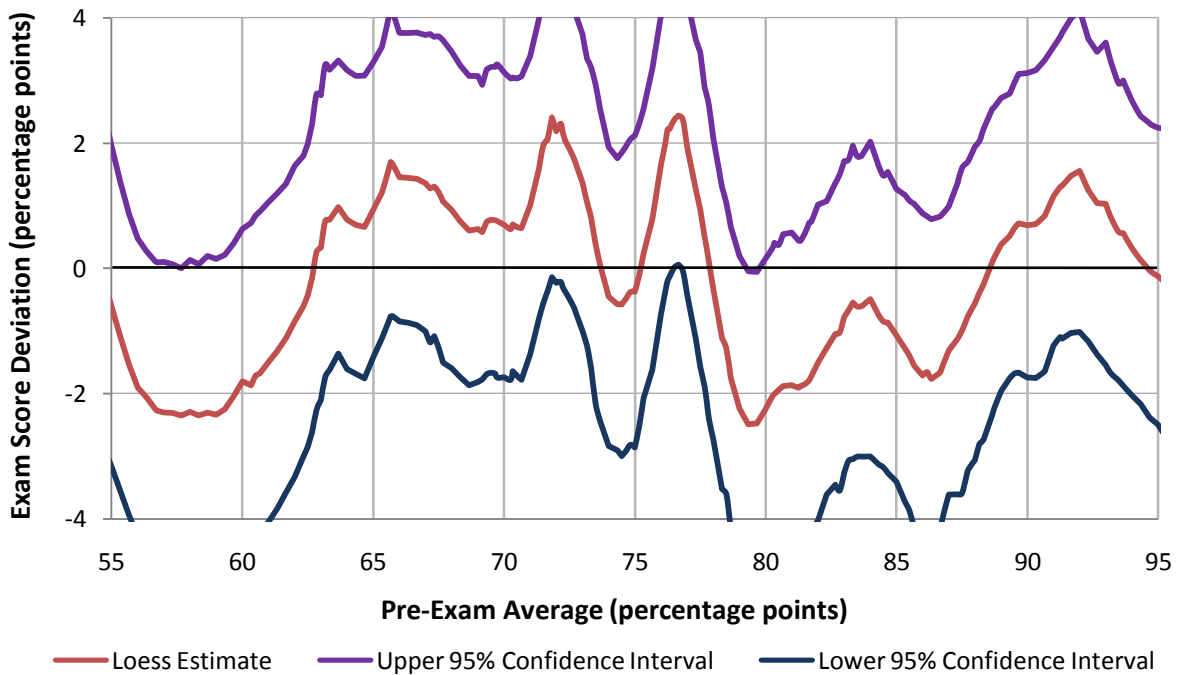
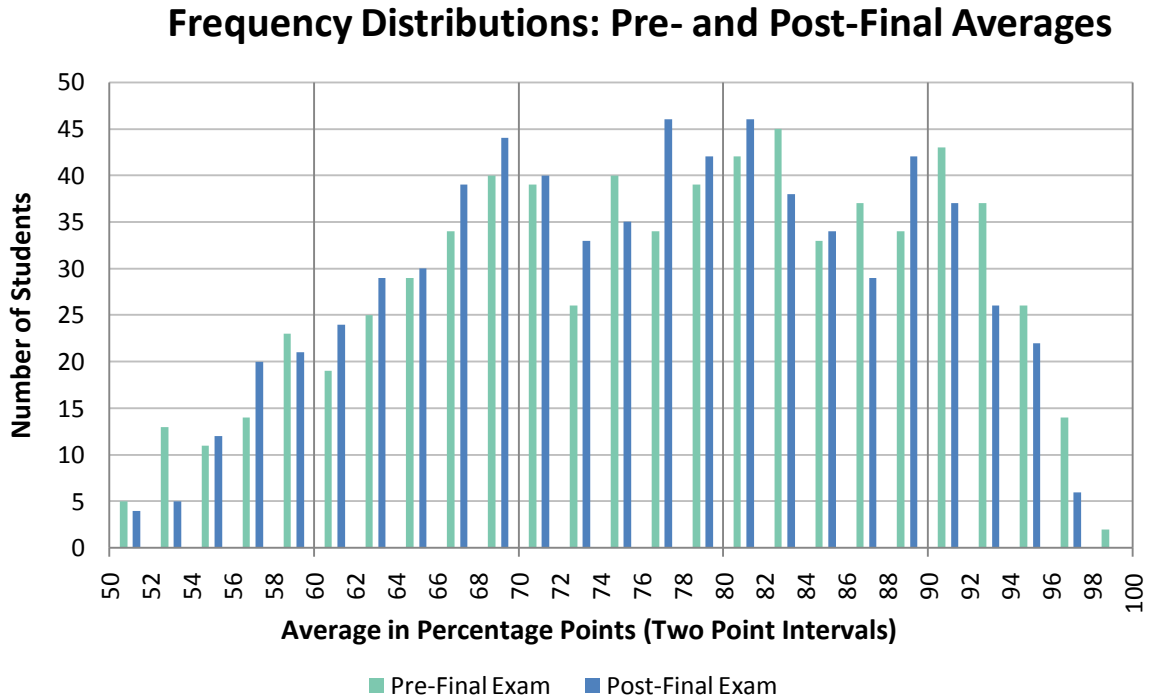


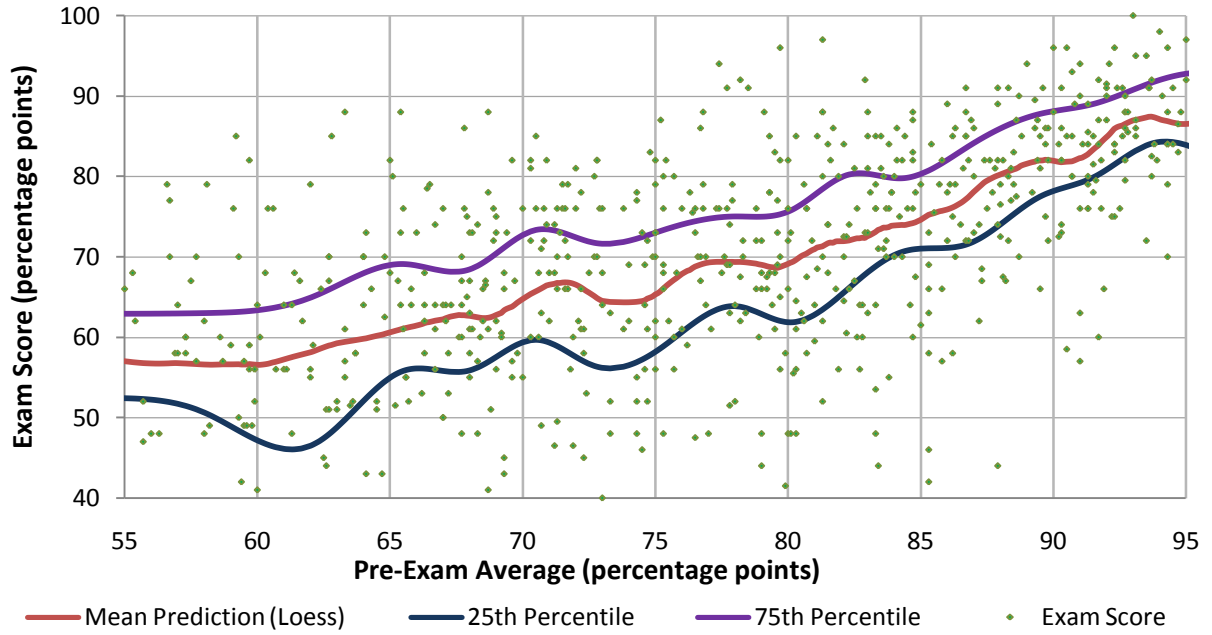
Figure 7. Results Portfolio: Hegwood.



TRANSITION MATRIX

Post-Final → Pre-Final ↓	Bottom Two Points of Range	Middle Six Points of Range	Upper Two Points of Range	Row Totals
Bottom 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)
<i>Column Totals</i>	<u>151</u> <u>(0.21)</u>	404 (0.57)	<u>149</u> <u>(0.21)</u>	704 (1.00)

Actual and Predicted Final Exam Scores



Exam Score: Deviation from Trend

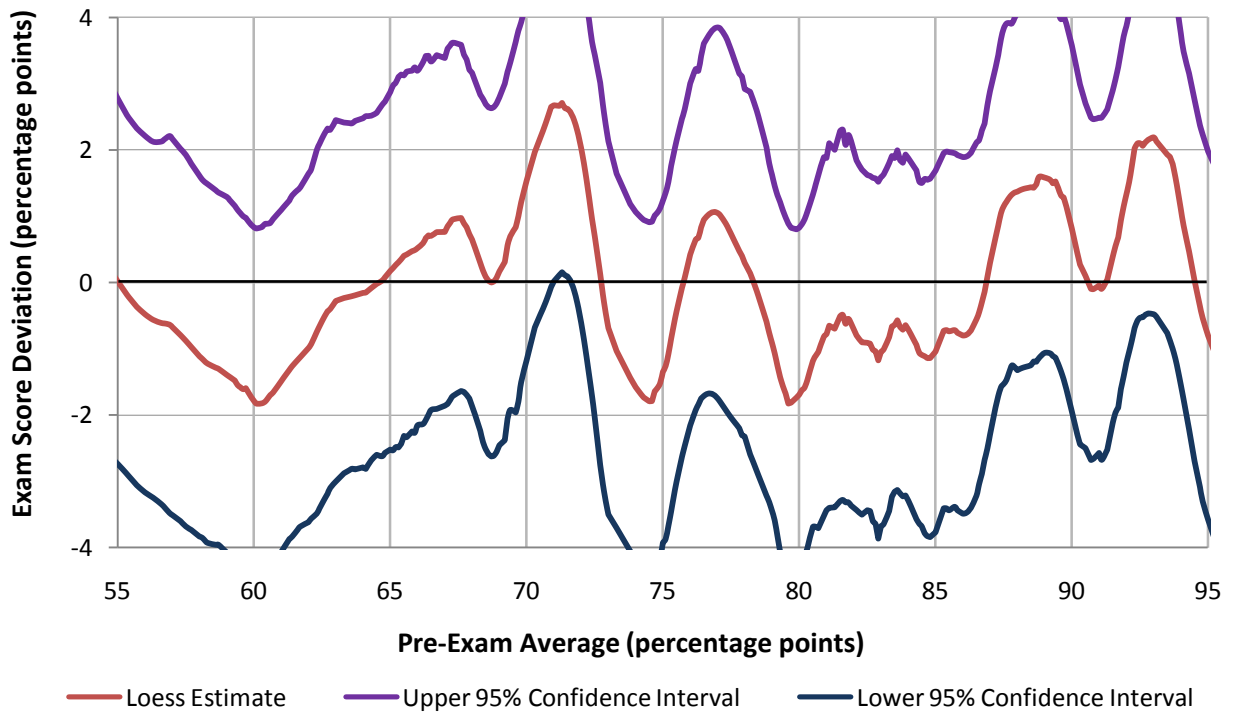


Figure 8. Abbreviated Results Portfolio: Sweeney.

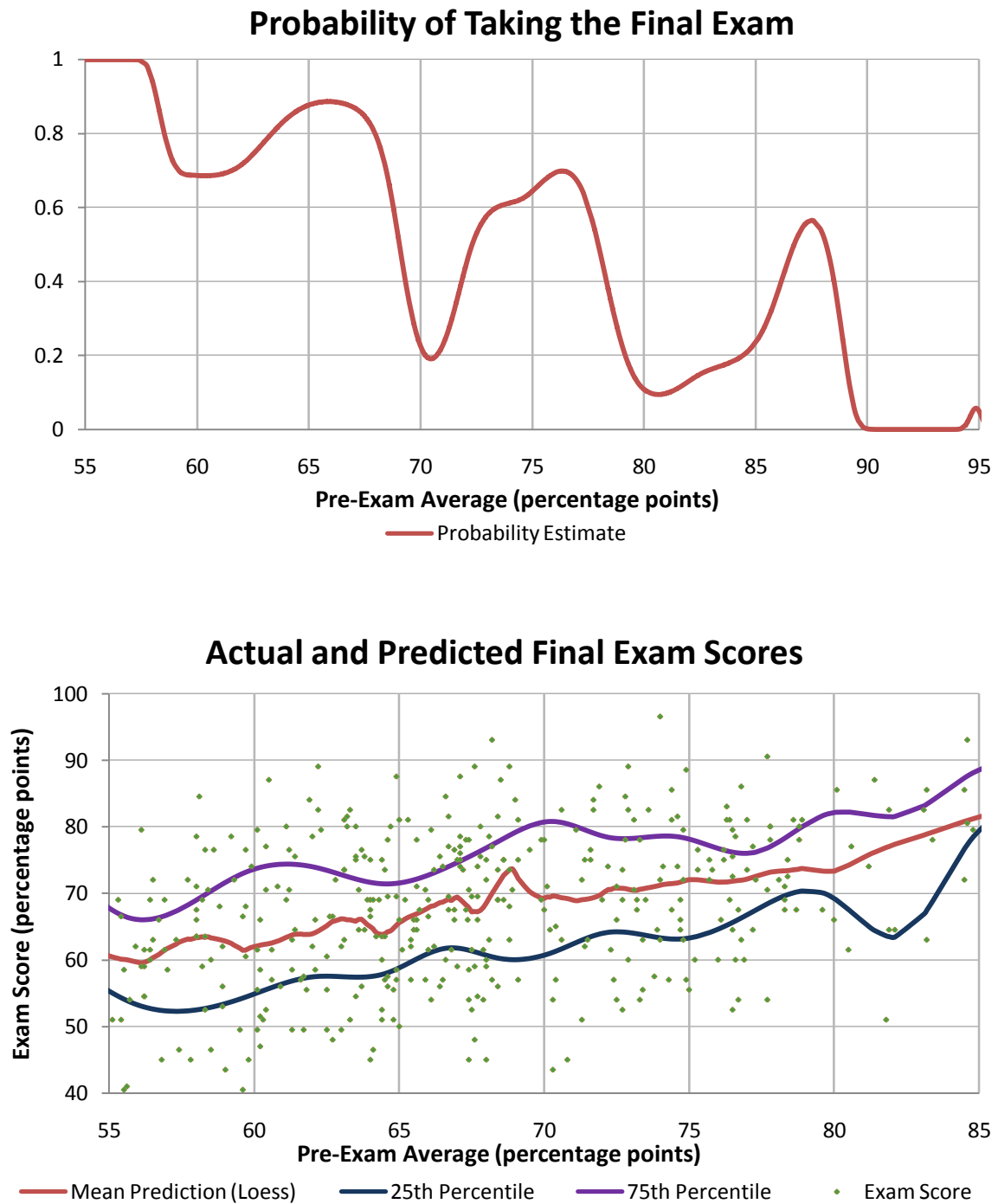


Table 1. Simulation Results. In the top table, the numbers in each cell are as follows: efficient effort, maximum effort provided under threshold, average effort provided under threshold, effort provided under direct measurement. In the bottom table, the numbers in each cell are as follows: the standard deviation of t among passers (top), failers (middle), and conditional on T (bottom).

$\gamma = 0.175$	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$	$\sigma = 5$
P = 1	1.013	2.083	3.017	3.960	4.844
	4.709	0.748	-1.569	-3.213	-4.488
	0.742	0.131	0.000	0.000	0.000
	0.406	-0.036	-0.972	-1.914	-2.793
P = 2	5.101	6.622	7.343	7.951	8.805
	8.670	4.709	2.392	0.748	-0.527
	1.617	1.570	1.008	0.258	0.000
	4.494	4.503	3.355	2.077	1.168
P = 4	12.949	14.008	14.665	14.555	14.487
	12.630	8.670	6.353	4.709	3.434
	2.256	2.832	2.976	2.737	2.262
	12.342	11.890	10.676	8.681	6.851
P = 7	24.932	23.877	22.846	21.940	21.201
	15.828	11.867	9.550	7.907	6.631
	2.517	3.496	4.081	4.307	4.255
	24.324	21.758	18.857	16.066	13.565
P = 10	27.489	28.862	27.627	26.123	24.963
	17.866	13.906	11.589	9.945	8.670
	2.638	3.821	4.639	5.100	5.272
	26.881	26.743	23.638	20.249	17.326

$\gamma = 0.175$	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$	$\sigma = 5$
P = 1	1.542	2.110	2.464	2.621	2.723
	2.633	2.355	2.464	2.621	2.723
	0.948	1.662	2.116	2.392	2.563
P = 2	1.304	1.338	1.823	2.454	2.723
	3.742	3.531	3.123	2.783	2.723
	0.948	1.662	2.116	2.392	2.563
P = 4	1.148	0.994	1.153	1.492	1.876
	3.849	3.571	3.233	3.098	3.017
	0.948	1.662	2.116	2.392	2.563
P = 7	1.054	0.820	0.924	1.189	1.491
	0.227	1.624	1.848	1.944	2.067
	0.948	1.662	2.116	2.392	2.563
P = 10	1.002	0.731	0.822	1.070	1.347
	0.211	0.504	0.829	1.167	1.493
	0.948	1.662	2.116	2.392	2.563

Table 2. Course Characteristics and Descriptive Statistics.

	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	2-3 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
Time Period for Data	2002-2007	2004-2007	1998-2007	2002-2007	2005-2007

Table 3. Parametric Estimates (coefficient estimates, with standard errors in parentheses).

	Pre-exam Average (a)	$(a/100)^2$	$(a/100)^3$	$1 \leq \delta < 2$	$2 \leq \delta < 3$	$3 \leq \delta < 4$	$4 \leq \delta$	Joint Sig. Test Stat. (p value)
<i>Berg</i>								
OLS	0.87 (0.34)	-2.49 (22.8)	----	----	----	----	----	----
OLS	3.39 (2.95)	-346 (407)	153 (185)	-0.97 (1.10)	-0.59 (1.11)	-0.72 (1.10)	0.08 (1.06)	0.38 (0.82)
LAD	5.29 (2.97)	-589 (395)	258 (173)	-1.62 (1.46)	-0.48 (0.88)	-1.52 (1.11)	-0.97 (0.91)	3.30 (0.51)
<i>Grant</i>								
OLS	-0.53 (0.58)	82.9 (37.6)	----	----	----	----	----	----
OLS	-2.78 (5.22)	361 (693)	-112 (303)	-0.16 (1.74)	-1.67 (1.83)	-1.10 (1.81)	-2.66 (1.69)	0.84 (0.50)
LAD	-1.93 (7.31)	240 (963)	-54 (418)	-0.63 (2.22)	-3.68 (2.03)	-2.94 (2.61)	-4.04 (2.27)	5.25 (0.26)
<i>Green</i>								
OLS	1.58 (0.38)	-64.9 (24.7)	----	----	----	----	----	----
OLS	5.19 (3.47)	-553 (466)	216 (206)	0.28 (1.21)	1.42 (1.20)	1.20 (1.21)	0.44 (1.14)	0.52 (0.72)
LAD	5.20 (3.51)	-547 (463)	211 (201)	0.49 (1.24)	0.03 (1.53)	0.23 (1.24)	-0.58 (1.21)	0.90 (0.92)
<i>Hegwood</i>								
OLS	-1.29 (0.42)	136 (27.6)	----	----	----	----	----	----
OLS	-2.49 (3.75)	284 (506)	-59 (225)	0.07 (1.24)	1.30 (1.26)	1.19 (1.30)	-0.95 (1.22)	1.13 (0.34)
LAD	1.40 (3.57)	-262 (484)	191 (215)	-0.42 (1.49)	0.35 (1.32)	0.58 (1.32)	-0.35 (1.34)	0.36 (0.99)
<i>Oettinger</i>								
OLS	-1.24 (1.11)	224 (156)	-70 (75)	-0.9 (0.7)	-0.5 (0.8)	-0.5 (0.7)	-1.9 (0.7)	(0.41)
LAD	-0.98 (0.80)	200 (125)	-62 (63)	-0.9 (0.9)	-1.0 (0.9)	-1.4 (0.9)	-2.7 (0.9)	(0.05)

Note: The test statistic is the F-statistic in the OLS regressions, and the chi-squared statistic in a likelihood ratio test in the Least Absolute Distance regressions. The dependent variable is the student's final exam score, in percentage points. Sample sizes are found in Table 1. δ is the absolute distance from a grade threshold, in percentage points. Semester dummies also included.