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

The Simple Economics of Thresholds: Evidence from the Western States 100

Darren Grant

SHSU Economics & Intl. Business Working Paper No. 10-04
November 2010

Abstract:

Many public and private entities utilize incentive systems in which improvements in measured performance are rewarded only if the agent crosses some pre-specified threshold. But neither the theory of their incentive effects nor the methods of estimating them has been fully developed. This paper comprehensively analyzes thresholds' positive and normative properties, lays out a simple and natural empirical strategy for estimating their incentive effects, and presents multiple applications of both. The strongest effects are exhibited by ultramarathoners trying to complete a one hundred mile race in under twenty-four hours.

THE SIMPLE ECONOMICS OF THRESHOLDS: EVIDENCE FROM THE WESTERN STATES 100*

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

Abstract: Many public and private entities utilize incentive systems in which improvements in measured performance are rewarded only if the agent crosses some pre-specified threshold. But neither the theory of their incentive effects nor the methods of estimating them has been fully developed. This paper comprehensively analyzes thresholds' positive and normative properties, lays out a simple and natural empirical strategy for estimating their incentive effects, and presents multiple applications of both. The strongest effects are exhibited by ultramarathoners trying to complete a one hundred mile race in under twenty-four hours.

JEL Codes: L83, C14, D10

Keywords: thresholds; behavioral incentives; ultramarathons

Word Count: 8,300

*** This paper has several figures that are best viewed in color. ***

* This paper belongs to a trilogy on the economics of thresholds. In the application here, the threshold has strong incentive effects, which confirm the properties derived herein. In the application in the companion paper, Grant and Green (2010), there are no incentive effects. A related paper, Grant (2010), sketches out thresholds' incentive effects when there is perfect measurement and population heterogeneity in the structural preference parameters. Sohna Jaye and Mitchell Graff helped gather and process the data, for which I am grateful. Curtis Barton guided me to the application presented in this paper.

Public and private entities frequently measure and reward 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. This simple change to the reward structure dramatically affects its incentive properties. With perfect performance measurement, the marginal benefit of improved performance is nil unless one crosses the threshold. With imperfect measurement, so that passing the threshold is uncertain (conditional on performance), expected marginal benefits are nonmonotonic, rising and falling rapidly in the neighborhood of the threshold. Both are 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. While thresholds such as these are a common feature of economic life, however, their positive and normative properties have not been fully developed. Thus, a unified discussion of these properties is warranted, along with a concordant, comprehensive estimation strategy.

In this paper we offer such a development, introducing a basic model of thresholds that generates a sequence of robust predictions that can be applied to a wide range of economic activity and tested elegantly with simple nonparametric methods. This advances the literature in three ways:

- 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, comprehensive, and revealing econometric strategy is introduced.

Two direct applications of these results are then provided: the strongest example of threshold incentive effects that we have been able to find, in, ironically, an ultramarathon, and an example of

an impotent threshold, excerpted from the companion paper, Grant and Green (2010). Finally, we show how these methods could be applied to several papers on the economics of education, sometimes to strengthen the authors' empirical tests, sometimes to overturn their conclusions.

I. The Behavioral and Normative Effects of Thresholds.

Behavioral Effects. 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).

But measurement exhibits diminishing returns, so it may be impractical to measure t precisely. This is true in a wide variety of circumstances, including many of those listed in Table 1 and most of the empirical applications discussed below. 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 $\mathbf{p}\sigma_\nu^2/(\sigma_\nu^2+\sigma_\epsilon^2) < \mathbf{p}$,[1] so each individual underprovides effort. The information provided to the market and the effort

[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.

elicited by agents can be improved, and under the right circumstances thresholds can do this. *Thresholds can be justified by imperfect information.*

Let the evaluator establish a passing threshold normalized, for simplicity, to 0. Instead of releasing T it simply indicates whether or not $T \geq 0$. The market value of passing the threshold is $\mathbf{P} = (\bar{t}_{PASSERS} - \bar{t}_{NONPASSERS})\mathbf{p}$; the probability of passing the threshold, conditional on effort, is now $\Phi((\nu+f)/\sigma_\epsilon)$, where Φ is the standard normal distribution function. 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 (as sketched out by Becker and Rosen, 1992). These are easily calculated and depicted when the costs of effort are specified as $C(f) = k\bullet(\exp(\gamma f) - 1)$, with k normalized to one and $\gamma>0$ representing diminishing returns or fatigue in the provision of effort. Now the logged marginal expected returns to effort, $\log(\mathbf{P}\Phi')$, form a quadratic in f, while the log of marginal costs, $\log(\gamma\bullet\exp(\gamma f))$, are a line.

Under these assumptions, it is easy to calculate effort among those agents who try at all:

$$f(\nu) = -(\gamma\sigma_\epsilon^2 + \nu) + \sigma_\epsilon\sqrt{\gamma^2\sigma_\epsilon^2 + 2\gamma\nu + 2\ln(0.4\mathbf{P}/\gamma\sigma_\epsilon)} \qquad \textbf{(1)}$$

The awkwardness of this expression belies the simplicity of the result: this equation represents (a segment of) a parabola with an axis of symmetry that runs through the origin and has a slope of -1. This is natural because the solution for $f(\nu)$ is the intersection of linear marginal costs and quadratic expected marginal benefits. Depicting this graphically is useful for generating heuristics.

Accordingly, Figure 1 represents five agents, A-E, whose upward sloping marginal cost of effort lines begin at ν_A-ν_E. For sufficiently low ν, as for agent A, marginal costs and marginal benefits do not intersect, so f=0: it is too much work to try to pass the threshold. This can also be true when

the curves do intersect, as the maximum may only be local, as between agents A and B, where total benefits are less than total costs. This continues until one reaches the extensive margin, where it is optimal to put forth effort (agent B). Effort then exhibits a discontinuity and becomes positive.

Clearly, this margin is always reached where $\nu < 0$. It may be also reached where $t < 0$, as in the figure; if so effort increases until it reaches its maximum, for agent C, at the vertex of the parabola, and declines steadily thereafter (agent D) until, at sufficiently high, positive ν, it returns to nil (agent 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.

Figure 2 depicts the resulting $\{\nu,f\}$ and $\{\nu,t\}$ loci for the non-trivial situation in which some agents put forth effort. The relation between natural ability and effort exhibits five properties, depicted in the figure and described heuristically below, with proofs sketched in a footnote.[2]

1. **Peak Effort Property:** *Colloquially, 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 property stems from the non-monotonic returns to effort. The existence of a point of peak effort (though not its location) has been previously shown (Oettinger, 2002, and others).

---

[2] Four properties of $f(\nu)$, when positive, are as follows: P1) $f' > -1$; P2) $f'' < 0$; P3) $f''' > 0$; and P4) $f' = 0$ implies $f = -\nu$ and $t = 0$. P4 (along with P1, when the maximum is at the extensive margin) ensure $\max(f) \geq -\text{argmax}\, f(\nu)$, so that $t(\text{argmax}\, f(\nu)) \geq 0$. These individuals' chances of passing the threshold are at least 50%, proving the Peak Proximity Property. And $\max(f) \geq -\text{argmax}\, f(\nu)$, along with P1, ensures $f(0) > 0$, the Precautionary Effort Property.

The Sawtooth Property is trivial if $\nu^* = \text{argmax}\, f(\nu)$ occurs at the extensive margin. For interior maxima, along with the extensive margin, $f'(\nu^*-d) > -f'(\nu^*+d)$ for any $d > 0$ by $f' = \int f''$ and P3. This property, along with P1 and $t = f + \nu$, ensures the Stair Step Property.

Because $f=0$ below the extensive margin, while $f(0) > 0$, P2 and P4 ensure that $f(\nu)$ has a single peak for some $\nu^* < 0$, possibly at the extensive margin. Thus one can define a region of $\nu < 0$ for which effort is higher than anywhere else: the Peak Effort Property.

2. **Sawtooth Property:** *Effort rises more quickly than it falls; that is, line BC in Figure 2 (top) rises faster than line CE falls, so that the {ν,f} locus takes a sawtooth shape.* This follows both from the existence of the extensive margin, at which effort increases discretely, and from the geometry of Figure 1. The point of intersection responds more to increases in the pre-exam average when marginal costs and expected marginal benefits are more similarly sloped, which occurs to the left of point C.

3. **Peak Proximity Property:** *Line OC in Figure 2 (top) has a slope ≤ -1, so that those individuals who try the hardest–whose ability is argmax f(ν)–have at least a 50% chance of passing the threshold.* This is a natural consequence of increasing returns to effort for $t < 0$.

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.

5. **Stair Step Property:** *More able individuals 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 {ν,t} locus always slopes upward, fastest near the extensive margin, like the sloping stair step at the bottom of Figure 2.

These predictions are fairly general, and can be supported with geometric arguments that do not depend on our specific functional forms. Furthermore, all extend to a broader interpretation of ν and f, in which the former represents a combination of ability and "base" effort and the latter represents the "strategic" effort perturbation in response to threshold incentives. Clearly this interpretation should prevail in our application to ultramarathons, which cannot be completed without tremendous effort. Still, the threshold incentive may spur the provision of even greater effort.

Normative Properties of Thresholds. We can now examine three reasons a threshold might be actively preferred to a system of direct measurement.

Motivating. Effort is underprovided under direct, imperfect 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, or perfect measurement, level of effort is small. Given our assumed functional forms, for example, one can show $f^*_{IMPERFECT} = (1/\gamma)[\ln(\mathbf{p}/\gamma) + \ln(\sigma_\nu^2/(\sigma_\nu^2+\sigma_\epsilon^2))] = f^*_{PERFECT} - (1/\gamma)\ln(1+\sigma_\epsilon^2/\sigma_\nu^2)$, so that $f^*_{IMPERFECT}/f^*_{PERFECT} = 1 - \ln(1+\sigma_\epsilon^2/\sigma_\nu^2)/\ln(\mathbf{p}/\gamma)$, which approaches zero as **p** declines.

Under these circumstances, thresholds can improve efficiency by intensifying the effort of individuals near the threshold. The rewards for passing, $\mathbf{P} = (\bar{t}_{PASSERS} - \bar{t}_{NONPASSERS})\mathbf{p}$, 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. It is not difficult to construct examples where effort is increased by the use of thresholds, just as one can construct examples where bundling, as a price discrimination mechanism, increases profits, or where competition for patents leads to "premature applications of discoveries" (Barzel, 1968). All work on the same principle: they convert a problem of the *intensive* margin, of marginal analysis, into a problem of the *extensive* margin, of whether participation in the activity (providing effort, purchasing the product, investing in innovation) is worthwhile. This additional margin provides leverage that can be used to increase effort.

Three such examples, from simulations of our model, are provided in panels A, C, and D of Figure 3.[3] In panel C, agents apply effort only under the threshold, not under direct measurement. But examples in which the threshold increases *efficiency*, not just effort, are more elusive. These generally require a high degree of imprecision in measurement ($\sigma_\epsilon \approx \sigma_\nu$), because this is when effort

[3] The word "simulation" is almost too strong: these are simply numerical calculations of the function $f(\nu)$ past the extensive margin, for the parameter values indicated. The figure identifies the values of the parameter **P**. Given this value and mean t for passers and nonpassers, **p** is backed out and used in calculating $f^*_{IMPERFECT}$ and $f^*_{PERFECT}$ under direct performance measurement. These effort levels do not depend on ν. The term $f_{THRESHOLD}$ refers to mean effort across all agents, in the presence of the threshold. All calculations are conducted for all values of ν, not just those shown in the figure.

under direct measurement falls far short of efficiency. In practice this situation is not only unlikely, but also inauspicious for employing thresholds, as passing or failing will be substantively due to luck, generating fairness concerns. Thus efficiency increases only in panel A, in which σ_ϵ is large.

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 discard valuable information on which to condition. Thresholds are never optimal for signaling.

Measuring Performance. 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 variance of t for passers, $\text{var}(t \mid T>0)$, is less than $\text{var}(t \mid T)$ when performance is measured directly.

Simulations not reported here can easily produce such an outcome, particularly with high 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 potential for gaming in this system is now well recognized (for example, Bolton, Freixas, and Shapiro, 2009), one can also see how within-grade risk clusters together, potentially enhancing the informational value of discrete ratings.

We now have two possible explanations for using thresholds, which both rely on imprecision in performance measurement. When the incentives for effort are relatively weak, a threshold can augment effort and thus improve efficiency; when they are relatively strong, a threshold can improve the accuracy of performance information that is provided to the public.

II. Estimation.

Estimating Behavioral Effects. To estimate thresholds' incentive effects, one must relate T to v using micro data, employing a specification that allows agents near the threshold to exhibit unusually strong performance. Four different regression approaches can be used to do this, but only one, nonparametric regression, is well suited to the essential econometric task, because the location and shape of the effort perturbation induced by thresholds cannot and should not be pre-specified. It is thus best identified by flexible methods, which also facilitate the formal and informal testing of the properties listed above.

Before proceeding it is important to identify the scale of the incentive effects we expect to uncover. We believe thresholds will be typically employed where the expected incentive effect

(among those who try at all) can exceed the error in measurement, σ_ϵ, but is much smaller than σ_v, the variation in ability: $\sigma_\epsilon < \max f < \sigma_v$. This appears to occur regularly in practice, based on the findings of the many empirical studies in Table 1 and discussed below. It also satisfies the fairness concerns articulated above. It may be difficult to identify threshold effects if $\max f > \sigma_v$.

We describe our preferred estimation methodology after critically examining two alternatives.

Regression Discontinuity. This technique may seem natural because thresholds induce an effort discontinuity at the extensive margin. But these methods were designed for situations in which the discontinuity is *imposed* by the "experimenter" and the size of the discontinuity identifies the effect of the intervention. So, for example, Card, Dobkin, and Maestas (2008) use an age discontinuity to estimate the effect of Medicare coverage, for which Americans qualify on their sixty-fifth birthday, on health care utilization.

But these conditions do not apply to the problem studied here. First, the discontinuity arises endogenously through optimizing behavior, which means its location, somewhere below the threshold, is not known in advance. Second, the discontinuity does not fully describe thresholds' incentive effects, which may or may not be largest at the extensive margin, and which extend even to individuals above the threshold ($\nu > 0$). Finally, when estimating mean incentive effects within a population, the discontinuity strictly arises only if all agents share common values of all structural parameters, which is unlikely in practice, as we will soon demonstrate. The underlying assumption for estimation purposes must be that these parameter values are sufficiently similar that the properties outlined above, which do not depend on a strict discontinuity, continue to obtain. For all of these reasons, regression discontinuity methods are impractical.

"Reduced Form" Parametric Estimation. Another alternative is to conduct a parametric

regression of T on v and include one or more dummy variables to capture perturbations in performance near the threshold, as follows:

$$T = \alpha + \beta v + \sum_j \rho_j 1(a_j \le v < a_{j+1}) + \lambda X + \xi \qquad (2)$$

where the values of a, specified in advance, define the ranges of v represented by each dummy, and X contains control variables. Perturbations in performance are identified by the ρ coefficient estimates, whose joint significance can be tested statistically (as in Oettinger, 2002).

This approach, while reasonable, is not optimal. If the dummies are few in number, it may be difficult to confirm the properties listed above; if they are too numerous, the coefficient estimates will be suffused with noise. In a way, this parametric approach is both too smooth and not smooth enough. The effect is required to be the same within the ranges specified by the dummy variables (by the a sequence), but is allowed to vary greatly across those ranges. The natural solution to this problem is nonparametric estimation, which allows a smooth, unrestricted estimate of the incentive effect across the entire domain.

Nonparametric Estimation. In this technique the $\{v,f\}$ or $\{v,t\}$ loci are estimated directly, leading to empirical results in the format of Figure 2. Unlike the alternatives, nothing need be pre-specified–not the intervals spanned by dummy variables, as in the parametric model, nor a discontinuity, as in the regression discontinuity model, nor a functional form, as in the structural model discussed below. (The following discussion relies on the survey by Yatchew, 1998.)

Furthermore, every feature of this problem is well suited to the characteristics of the data. There is typically just one independent variable of interest, v or its proxy, which eliminates the "curse of dimensionality" and allows specification tests that use simple differencing methods. (Control

variables can be included, but this can probably be done parametrically, holding the curse of dimensionality at bay.) Estimation is conducted on individual microdata, which is often numerous enough to permit reasonably precise nonparametric estimates, in which the effect of the threshold on outcomes can be directly determined. And one can informally determine whether these estimates are consistent with the five properties listed above, by identifying the empirical equivalents of point C, point D, line BC, and line CE in Figure 2, and comparing their values, slopes, or relative slopes to those predicted by the theory.

Most importantly, a simple, natural, formal test for threshold incentive effects can be conducted–not a parameter test, but a more comprehensive, more powerful specification test. In the absence of threshold incentive effects, outcomes should be a smooth function of natural ability. In our theoretical model, for example, there is a direct linear relationship. Allowing the units of measurement of T and ν to differ, and allowing there to be control variables X, this linear relationship is represented by this parametric regression:

$$T = \alpha + \beta\nu + \lambda X + \xi \qquad \textbf{(3)}$$

The adequacy of this regression is the null hypothesis. The alternative is that this parametric relation is inadequate, because effort is systematically related to proximity to the threshold. Absent controls, a very simple specification test is based on A, the average squared error in equation **(3)**, and B, one half of the mean squared difference between adjacent values of T, after being placed in ν-order:

$$Z = \sqrt{S}\cdot(A - B)/B \sim N(0,1) \qquad \textbf{(4)}$$

where S is the number of observations. The null is rejected for sufficiently large values of the test

statistic *Z*. Yatchew (1998) provides other practical alternatives.

If the null is rejected, and one can pre-specify an ability range over which threshold incentive effects will not appear, $\nu < \nu_L$ or $\nu > \nu_H$, the effort perturbation $g(\nu)$ can be estimated as follows:

$$T = \alpha + \beta\nu + g(\nu)\cdot 1(\nu_L < \nu < \nu_H) + \lambda X + \xi \qquad \textbf{(5)}$$

These perturbations should satisfy the properties identified in Section I.

Structural Estimation and the Identification of Normative Effects. So far, we have deemed the least structured empirical approach best. Following this logic in the other direction, we also argue against using the most structured empirical approach, structural estimation, which would use the following regression specification:

$$T = \alpha + \beta\nu + f(\nu;\gamma,\sigma_\epsilon,\boldsymbol{P}) + \lambda X + \xi \qquad \textbf{(6)}$$

While the function *f* takes a simple form, given in equation **(1)**, the domain over which it is positive (that is, the location of the extensive margin) must be calculated numerically for each $\{\gamma,\sigma_\epsilon,\mathbf{P}\}$ combination. Estimation, while feasible, is difficult. Given estimates of **P** and σ_ϵ and the distribution of *T*, the final structural parameter, **p**, can be calculated, but this also requires numerical techniques.

Our primary objection to this approach, however, is based not on its feasibility, but its utility: the structural parameters cannot be cleanly resolved. Under the restrictions $\sigma_\epsilon < \max f < \sigma_\nu$ the shape of the $\{\nu,f\}$ profile is typically quite insensitive to the parameter values–and, thus, the parameter estimates are quite sensitive to the shape of the estimated profile. Scaling the horizontal axis so that $\sigma_\epsilon = 1$, **P** is the following function of the *f* (vertical) and positive ν (horizontal) intercepts:

$$\boldsymbol{P} = \frac{v_{int}^2 - f_{int}^2}{f_{int} - 1} \cdot e^{0.5 v_{int}^2} \qquad (7)$$

This equation *can* be very sensitive to changes in the value of the either intercept, and *will* be when the restrictions above obtain, as (one can show) the horizontal intercept will exceed one. (That is, agents one standard deviation of measurement error above the threshold will still try at least a little.)

To illustrate, the simulations in panels B, C, and D of Figure 3 illustrate three ability-effort profiles satisfying these restrictions. All three figures have very similar profiles, yet very different parameter values–and different normative implications. In panel B threshold effort is underprovided and inefficient, compared to direct measurement; in panel C it is overprovided and inefficient; in panel D, it is overprovided to roughly the same degree that effort under direct measurement is underprovided. Quantitative assessment of thresholds' efficiency properties is often impractical.[4]

---

[4] Under some circumstances, however, there may be practical alternatives. Sometimes qualitative judgements may be possible, as in the companion paper. Other times, the effort under direct measurement may be known. Finally, and surprisingly, the net incentive effect can be signed using just one structural parameter, σ_ϵ, which can sometimes be imputed a priori.

Too see this, recognize that threshold effort satisfies the following condition, which equates marginal costs and expected marginal benefits, in logarithms: $\gamma f_{\text{THRESHOLD}} = \ln(\mathbf{P}) - \ln(\gamma) + \ln(\phi(t))$. Using $\mathbf{P} = (\bar{t}_{\text{PASSERS}} - \bar{t}_{\text{NONPASSERS}})\mathbf{p} = \Delta\bar{t}\mathbf{p}$ and the results above, $\gamma f_{\text{THRESHOLD}} = \gamma f_{\text{IMPERFECT}} + \ln(\Delta\bar{t}) + \ln(1+\sigma_\epsilon^2/\sigma_\nu^2) + \ln(\phi(t))$. Simplifying the last term and rearranging yields: $\gamma(f_{\text{THRESHOLD}} - f_{\text{IMPERFECT}}) = \ln(\Delta\bar{t}) + \ln(1+\sigma_\epsilon^2/\sigma_\nu^2) - \ln(2.5\sigma_\epsilon) - t^2/2\sigma_\epsilon^2$. This expression depends on ν, which (naturally) determines t. Taking expectations across ν, and using the fact that $\text{var}(T) = \text{var}(t) + \sigma_\epsilon^2$, yields:

$$\gamma E(f_{\text{THRESHOLD}} - f_{\text{IMPERFECT}}) = -0.4 + \ln(\Delta\bar{t}) + \ln(1+\sigma_\epsilon^2/\sigma_\nu^2) - \ln(\sigma_\epsilon) - (\text{var}(T) + \bar{T}^2)/2\sigma_\epsilon^2.$$

The $\Delta\bar{t}$ and T terms can be inferred from the data; the others may be reasonably approximated. This equation is exact unless normality in ν, required to finesse the game-theoretic concerns noted in footnote 1, does not hold strictly, in which case it is approximate. This expression signs the difference in average effort under the two systems, but does not quantify it unless γ is known.

Distributional Analysis. 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 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.

III. Application: The Western States 100.

Because our purpose here is to demonstrate the ability of the techniques described here to reveal thresholds' incentive effects, our application has been chosen for its technical properties, not its social relevance. It is the strongest example of threshold incentive effects we have been able to find, has good proxies for "natural ability," and contains a large number of observations, facilitating nonparametric estimation. Two socially relevant examples from the author's other research, Grant (2010) and the companion paper, Grant and Green (2010), fail to demonstrate any threshold incentive effects. The latter is discussed below as a "counterexample," in which our techniques reveal the impotence of the incentive.

The Western States 100 (WS100) is one of the largest and most venerable ultramarathons in the U.S. (Its web site, www.ws100.com, contains most of the following background information.) Originating from a trail ride on horseback, the first official run was held in 1977 and quickly grew to

its current size of about 370 runners annually. The extremely challenging one hundred mile run, beginning in California's Sierra Nevada mountains and ending in the valley below, features repeated elevation changes, hot and cold temperature extremes, high altitudes, rugged trails, and night running.

Entry is primarily by lottery. The number of applicants equals about one thousand, and all accepted entrants must qualify by running reasonably good times in races of fifty miles or longer, or by completing a certified trail run of one hundred miles. Thus, the entrants in the race have shown the capacity to complete the WS100, but are not certain to complete it quickly.

The course closes after thirty hours, but a highly coveted medal is presented to all those finishing in under twenty-four hours, the time standard used for the original, equine ride. The winning time is approximately sixteen hours; about one hundred runners come in under twenty-four hours; about another two hundred finish between twenty-four and thirty hours, with the remainder dropping out of the race or finishing after the course has closed.

With few exceptions (such as years with wildfires), the WS100 has run the same course since its inception. Finish times and split times, for nine aid stations spread throughout the course, are recorded on the run's web site for the run's entire history. Eliminating years in which the course was changed or the location of the aid stations was moved (2003, 2002, 1998, 1995) and observations with incomplete split information yields a sample of 3,991 runners over the period 1986-2006. Split times are recorded in minutes, finish times in minutes and seconds.

Figure 4 illustrates the course layout. The course is effectively run in two stages. The first two-thirds of the course feature high elevations, steep gradients, and temperature extremes of mountain cold and daytime heat. (Even temperatures in the seventies are onerous in a race of this length.) Then, between the sixth and the seventh splits, the course drops to low altitude and flattens

out; day turns into night, and any daytime heat subsides. This begins the second stage.

As the stage changes, so do runners' racing strategies. Figure 5 illustrates the distribution of recorded split times in the full sample at splits 2, 4, 6, 7, 8, and the finish, using both a simple histogram and a more precise kernel density. During the first two-thirds of the race, split times take a bell-curve shape, but between splits 6 and 7 this distribution begins to bifurcate. This bifurcation grows until, at the finish, the density is, for practical purposes, divided into two highly skewed distributions: one bunched ahead of 1,440 minutes, or twenty-four hours, and another ahead of 1,800 minutes, the time that the course closes. This suggests contestants run the first stage of the race at a reasonably even pace, generating the ever-widening bell curves, and then tweak their times during the more manageable second stage to try to satisfy one of the two thresholds. The finish time kernel density indicates that many are, in fact, successful.

This finish time distribution is censored after 1,800 minutes, but the distribution surrounding the 1,440 minute threshold is not, and it unquestionably shows large threshold effects. A total of 97 runners finish no more than ten minutes ahead of the threshold; only 19 runners finish no more than ten minutes behind the threshold, a highly significant difference. We now determine whether these incentive effects are revealed in the nonparametric regressions sketched out above.

IV. Empirical Results.

To begin our empirical analysis, we need a proxy for "natural ability," and split times can serve this purpose. The previous discussion suggests a natural proxy: the time at the sixth split, just before the end of the "first stage" of the race. The top of Figure 6 presents a scatterplot of these split

times versus finish times, both measured in logarithms (which best fit the data), along with a (loess) smoothed estimate of the mean. Censoring, after which finish times are not recorded, occurs at the top of the graph at 7.50 log points. The medal threshold of 7.27 log points is easily visible as a horizontal "strip" of finish times: evidence, again, of a threshold effect.

This strip features "soft" horizontal and vertical edges, both of which are revealing. The soft horizontal edge confirms our assumption of measurement uncertainty. Without this we would expect points to be bunched not near the 7.27 line, but perfectly along it. Here, this uncertainty pertains not because the actual finish time is imprecisely measured, but because the runner cannot perfectly forecast his finish time while on the course, and cannot perfectly self-regulate his pace. In addition, if there were a single extensive margin for all runners, the right edge of this strip would terminate abruptly instead of steadily withering away, as in the figure.

The loess mean indicates that the relationship between the split time and the finish time is generally smooth, but does exhibit a slight perturbation near the threshold. Clearly, however, these threshold incentive effects are small relative to the overall variation in split and finish times, satisfying the scale restriction $\max f < \sigma_v$ adopted above. The smoothed mean always slopes upward, consistent with the Stair Step Property, but the stair step itself is obscured, because better times on this graph occurs to the threshold's left, not its right. To maintain the orientation used in Figure 2, Figure 6 should be held upside down–and then the stair step appears.

We begin with a parametric regression analysis, as in equation **(3)**. Figure 6 suggests that the sample should be restricted to those 2,273 individuals for whom the logged split time is less than 6.8, for their finish times are unlikely to be censored, and supports a simple linear relation between the split time and the finish time over this domain (as does an insignificant quadratic term, when included

in this regression). Year dummies are included as controls.

The residuals obtained from this regression are plotted in the next figure, Figure 7, for all but the fastest runners. To orient the figures in the same way the theory was presented, predicted finish times, on the horizontal axis, run from largest (worst) to smallest (best), as do the deviation of actual finish times from predicted, on the vertical axis. This axis, also measured in log points, is placed at the threshold, 7.27, and axis spans finish time deviations as large as 7%. Along with the mean residual, again calculated with a loess smoother, are 95% confidence intervals.

The relevance of threshold incentive effects is formally supported by the residuals-based specification test in equation **(4)**, which can be conducted although this regression contains year dummies, by finding the difference in finish times between adjacent runners (ordered by split times) running the WS100 in the same year. Our test statistic of 2.00 suffices to reject the null hypothesis of no misspecification. The figure indicates a perturbation in performance in the neighborhood of the threshold, and only in this neighborhood, that is associated with logged split times ranging from about 6.64 to 6.76. We structure the nonparametric term accordingly in the semiparametric regression, based on equation **(5)**, that is intended to identify this perturbation.

Figure 8 presents estimates of $g(v)$ in this regression, conducted with a loess smoother in SAS procedure GAM, with the bandwidth chosen by cross-validation and year dummies again included as controls. The $g(v)$ term is easily significant at $p < .01$, and indicates performance improvements of as much as 1.5%. The Peak Effort Property and the Precautionary Effort Property are transparent, while the Sawtooth Property is also supported (the acclivity is 50% steeper than the declivity). The Peak Proximity Property is also confirmed: at the point of maximum effort the runner has a three-fourths chance of passing the threshold. This, is turn, suggests that maximum effort occurs not at an

interior solution, as in Figure 2, but at the extensive margin, as in the profiles in Figure 3. Simulations suggest this virtually always occurs under the restriction $\sigma_\epsilon < \max f$. The gentle upward slope of the perturbation seems to contradict this finding, but we will soon reconcile this contradiction.

Extensions. A careful comparison of Figures 7 and 8 reveals that the smoothed residuals are almost identical to the perturbation estimate, except for an additive scale factor that arises because the grand mean of the residuals equals zero. This should not be surprising, because the parametric regression basically detrends the data, and the effort perturbation, being small in scale, affects the estimate of this trend only slightly. The same would be true of a more general trend, such as a smooth polynomial in v, which may be necessary if v is measured with heteroskedastic error, which will cause regression to the mean at a non-constant rate (as in Grant and Green, 2010).

In other instances, however, natural ability may be proxied by more than one variable. In the WS100, for instance, a stronger and sharper effort perturbation is observed in residuals from a parametric, double-log regression of the finish time on the sixth split time and the elapsed time between splits six and seven. The previous discussion implies that smoothed residuals from these regressions will also provide a good estimate of threshold incentive effects, but (with our nonparametric regression approach) a more formal option is also available: the single-index model. Ichimura's (1993) estimator is not yet available in commercial software, but Horowitz and Hardle's (1994) specification test can be conducted without estimating the full nonparametric model.

Alternatively, it may be valuable to estimate not the smoothed mean, but smoothed quantiles.[5]

[5] Nonparametric quantile estimation methods can be executed using the techniques discussed in Koenker (2005), or with commercial software using transformation regression, the method adopted here. This technique expresses the key independent variable v (here, the split time) as the sum of an

Here, this estimates the amount of incentivized effort among those runners who tend to "finish strong" and among those who don't. Estimates of the 25th, 50th, and 75th percentiles, presented in Figure 6 (bottom), are indeed revealing. Runners at the 25th percentile, who tend to run the second stage of the race relatively quickly, appear to increase their finish time by as much as 2% compared to trend, while those at the 75th percentile–who have more energy in reserve–increase their times by as much as 5%. The Stair Step and Sawtooth Properties are clearly followed in each instance, while the sudden and sharp onset of the effort perturbations, moving inward from the top right corner of the graph, indicate the locations of the extensive margins. The modest mean threshold effects that are illustrated in Figure 6 (top), Figure 7, and Figure 8 are in fact smoothed combinations of stronger threshold effects across runners at different quantiles.

V. Applications to Educational Research.

Thresholds are rife in education, where they delineate acceptable or noteworthy performance for students, educators, schools, and school districts. Increasingly, economic research utilizes these thresholds to draw conclusions about the effect of incentives on student, educator, and school performance. The results developed here can sharpen the conclusions drawn in some of these studies and modify the conclusions drawn in others. We illustrate with four varied examples.

---

overlapping series of "B-splines," $S_s(v)$, calculated using the method of deBoor (1978), such that $\Sigma S_s(v) = 1 \; \forall \; v$. These splines are then employed as independent variables in a parametric quantile regression. Applying the coefficient estimates to the splines yields a smooth, unrestricted estimate of that quantile. That is, if the coefficients on the splines are β_s, then the predicted value of the quantile at split time v is $\Sigma \hat{\beta}_s S_s(v)$. The extent of smoothing is governed by the number of splines.

Grant and Green (2010). How much are students motivated to study by the prospect of earning a higher grade? To answer this question, the companion paper applies the methodology developed here to micro data on student performance in four business courses taught by five instructors at two universities. Final exam performance was related to the proximity of students' pre-exam course averages to the threshold between two letter grades. Students near the threshold are expected to do unusually well on the final examination, compared to those far away.

While this might seem like a straightforward application of basic economic theory, this is not so: there is consistently no effect. The methodology here helps reinforce that counterintuitive result by its generality, showing that the failure to reject the null does not stem from an arbitrary parametric specification choice, and its completeness, showing that the Peak Effort Property and Sawtooth Property are repeatedly violated, while the Stair Step Property is occasionally violated.

Figure 9 presents a subset of results from that paper, which use data from this author's Principles of Microeconomics classes. The relation between students' course averages and their final exam scores is essentially a trend, uninterrupted by significant perturbations near the threshold dividing any two letter grades (A, B, C, D, F)–not even on the pass/fail border. A specification test fails to reject the null that this trend adequately describes the data. This is complemented by an analysis of the distribution of post-exam course averages, which are not bunched just above the cutoffs dividing grade thresholds: nineteen percent of all unrounded final course averages end with a units digit of 0 or 1, slightly below the 20% that would be expected at random. Furthermore, pre-exam/post-exam transitions in students' course averages reveal that, after taking the final exam, students are as likely to drop just below a grade threshold as to rise just above it. The italicized cells of the table, which contain these two transition probabilities, are not significantly different. All

together, in a total of twenty hypothesis tests conducted in that paper–five tests for each of four instructors–exactly one is significant at the five percent level and one other at the ten percent level, just as predicted by chance.

McEwan and Saltibañez (2005). This paper examines the effect of incentives on teacher effectiveness, using for identification a points threshold required for promoting schoolteachers in Mexico. One-fifth of the available points are generated by scores on standardized student tests, and McEwan and Saltibañez find that scores rise modestly just above a somewhat arbitrarily-chosen threshold that distinguishes teachers with a reasonable chance of being promoted from those who don't. They thus conclude these performance incentives are effective means of improving instruction.

These results are presented in Figure 10 (bottom). The positive effect, an increase in awarded points of about five percent above trend, is clearly visible. However, this finding violates the Peak Proximity Property: at the point of maximum effort, the chances of passing the threshold must be at least one-half. In the figure, however, these chances are below one-half: the total points received by the average individual at the point of peak effort, 53+11.5=63.5, are well below the 70 points required for promotion. Simply put, if it is worth expending great teaching effort if you "begin" with 53 points, it should be at least as valuable to do so at 54 points or more. Absent an explanation why this should not be the case, these empirical results must be viewed not as evidence of threshold incentive effects, but some unknown specification error that generates this spurious result.

Reback (2008). This paper looks at the effect of accountability standards in the state of Texas on the distribution of academic achievement. These standards classify schools into one of four

categories–low performing (unacceptable), academically acceptable, recognized, and exemplary–based (to simplify slightly) essentially on the fraction of students passing state assessments. Reback creates an index, called the "accountability incentive," that measures the "marginal benefit to the school from a moderate increase in a student's expected performance" (p. 1404). This incentive, depicted in Figure 10 (top) for the low performing/academically acceptable threshold, takes a shape resembling that in Figure 2, but this figure depicts the construction of the key *independent* variable. This variable is then used to predict test scores, and its significance indicates that schools focus their energies on those students closest to passing the threshold.

Our analysis suggests a less restrictive, more demanding way to explore the same topic: take the prediction of each students' expected assessment score that goes into forming the accountability incentive–which is v is our nomenclature–and nonparametrically relate it directly to actual test scores. If schools respond to assessment incentives as predicted, this regression would itself yield curves that look like those at the bottom of Figure 2, and residuals that resemble those in Figure 10, which would satisfy all the properties enumerated above. A closely related paper, Neal and Schanzenbach (2010) provides graphs that, while not formally testing the properties derived in Section I, appear to be consistent with several of them.

Craig, Imberman, and Perdue (2009), Ahn and Vigdor (2009). Using Texas and North Carolina data, respectively, these authors move further up the bureaucracy and explore how districts respond to accountability incentives, which typically assign schools one of three or four ratings based on the fraction of students that score acceptably on standardized tests.

Of course, these ratings are separated by thresholds (to simplify slightly, as the rating process

is not quite this simple). These thresholds can be used in one of two ways. Retrospectively, using regression discontinuity methods, one can compare consequences for schools that just did, and did not, achieve their desired rating. This is the approach used by both of these papers, which collectively find that successful schools receive a funding increase but perform no differently from their less successful confederates on future tests.

Alternatively, these thresholds could be used prospectively, to identify how threshold incentives affect schools' outcomes, using the methods developed here. Using the prior year's score as a proxy for "natural ability," one can discern whether there is particularly strong performance by schools near the thresholds separating rating categories, which satisfies the properties articulated above. This fundamental question about accountability incentives has so far gone unaddressed.

VI. Conclusions.

A basic economic model predicts several properties threshold incentive systems should possess. These properties can be easily checked, and the incentive effects revealed, using nonparametric regression. This practical method adds rigor and generality to the methods used heretofore, which can be applied to a wide range of phenomena.

REFERENCES

Ahn, Thomas, and Jacob Vigdor. "Does No Child Left Behind Have Teeth? Examining the Impact of Federal Accountability Sanctions in North Carolina," Manuscript, Duke University (2009).

Barzel, Yoram. "Optimal Timing of Innovations," *The Review of Economics and Statistics*, 50, 3:348-355 (1968).

Becker, William, and Sherwin Rosen. "The Learning Effect of Assessment and Evaluation in High School," *Economics of Education Review*, 11:107-118 (1992).

Bolton, Patrick, Xavier Freixas, and Joel Shapiro. "The Credit Ratings Game." Manuscript, Columbia University (2009).

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

Card, David, Carlos Dobkin, and Nicole Maestas. "The Impact of Nearly Universal Insurance Coverage on Health Care Utilization: Evidence from Medicare," *American Economic Review*, 98, 5:2242-2258 (2008).

Card, David, and Alan Krueger. "Time-Series Minimum Wage Studies: A Meta-analysis," *American Economic Review* 85,2:238-243 (1995).

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).

de Boor, C. *A Practical Guide to Splines*. New York: Springer Verlag (1978).

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*, 48: 756-770 (2010).

Grant, Darren, and William B. Green. "Grades as Incentives," Manuscript, Sam Houston State University, 2010.

Grundfest, Joseph A., and Nadya Malenko. "Quadrophobia: Strategic Rounding of EPS Data," Rock Center for Corporate Governance at Stanford University Working Paper No. 65 (2009).

Horowitz, Joel, and Wolfgang Hardle. "Testing a Parametric Model against a Semiparametric Alternative," *Econometric Theory*, 10:821-848 (1994).

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

Ichimura, H. "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models," *Journal of Econometrics*, 58:71-120 (1993).

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).

Keonker, Roger. *Quantile Regression*. Cambridge: Cambridge University Press (2005).

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).

Neal, Derek, and Diane Whitmore Schanzenbach. "Left Behind by Design: Proficiency Counts and Test-Based Accountability," *Review of Economics and Statistics*, 92, 2:263-283 (2010).

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).

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

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

Yatchew, Adonis. "Nonparametric Regression Techniques in Economics," *Journal of Economic Literature*, 36:669-721 (1998).

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

Topic	Selected Studies	Threshold	Theory	Evidence
gaming of bonus systems or financial reporting requirements	Healy (1985), Courty and Marschke (2004), Grundfest and Malenko (2009)	annual cutoff for meeting quotas to qualify for bonuses, or the 0.5 cent cutoff to round up earnings per share	emphasizes potential adverse effects of thresholds	timing of reported output is adjusted to maximize bonuses; small accounting adjustments are made to nudge up earnings per share to the next cent
criminal behavior, drunk driving	Friedman and Sjostrom (1993), Iyengar (2008), Grant (2010)	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]"
effort by students, schoolteachers, schools, or districts	Oettinger (2002), McEwan and Saltibanez (2005), Reback (2008), and many others	letter grade cutoffs; "points" required for promotion, for passing a high-stakes test, or for a higher school rating	emphasizes the "Peak Effort Property" described below	see the extended discussion below, especially Section V
analyst / publication bias in several fields of social science	Card and Krueger (1998), 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

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

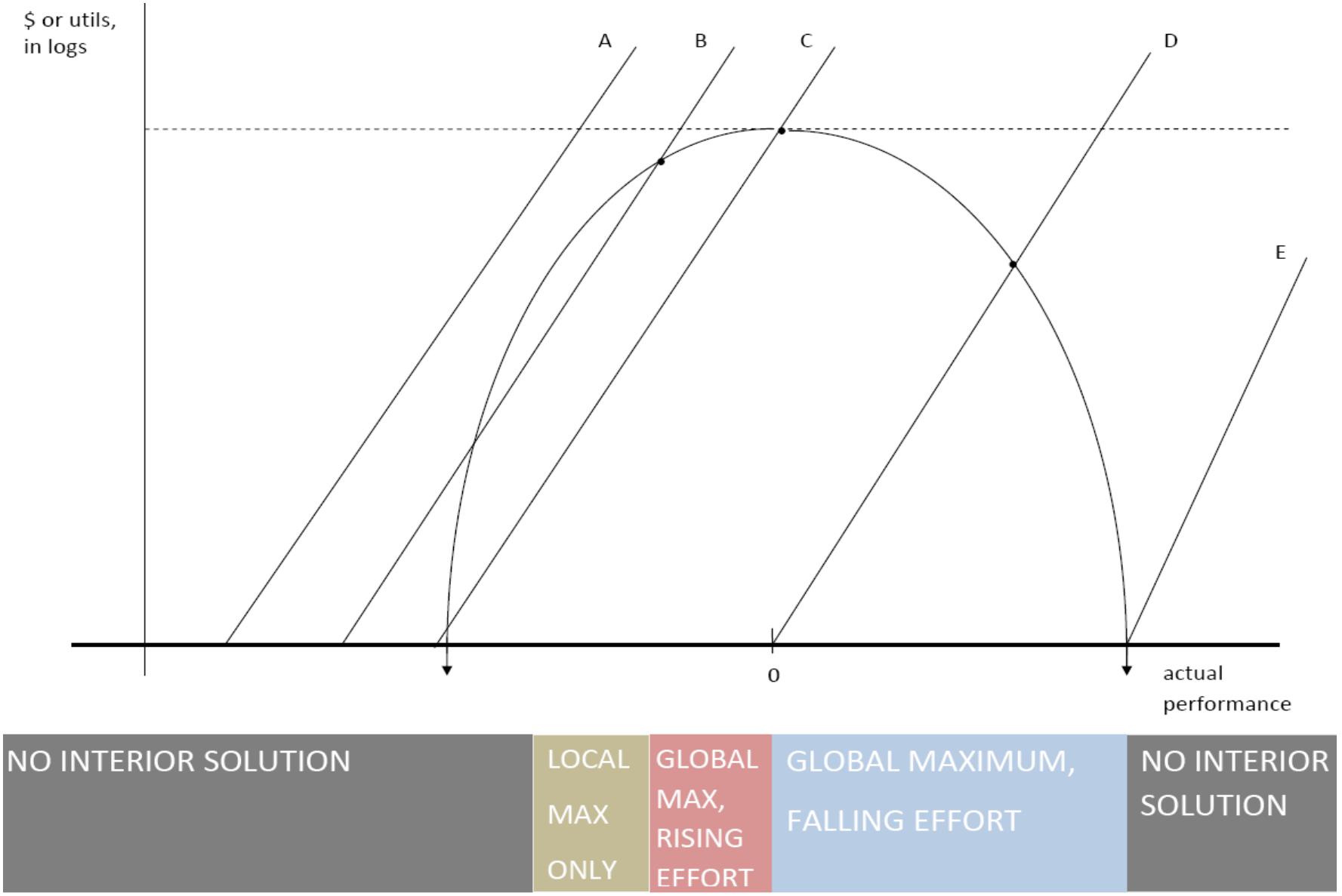


Figure 2. Top: Ability-Effort Locus. Bottom: Ability-Performance Locus.

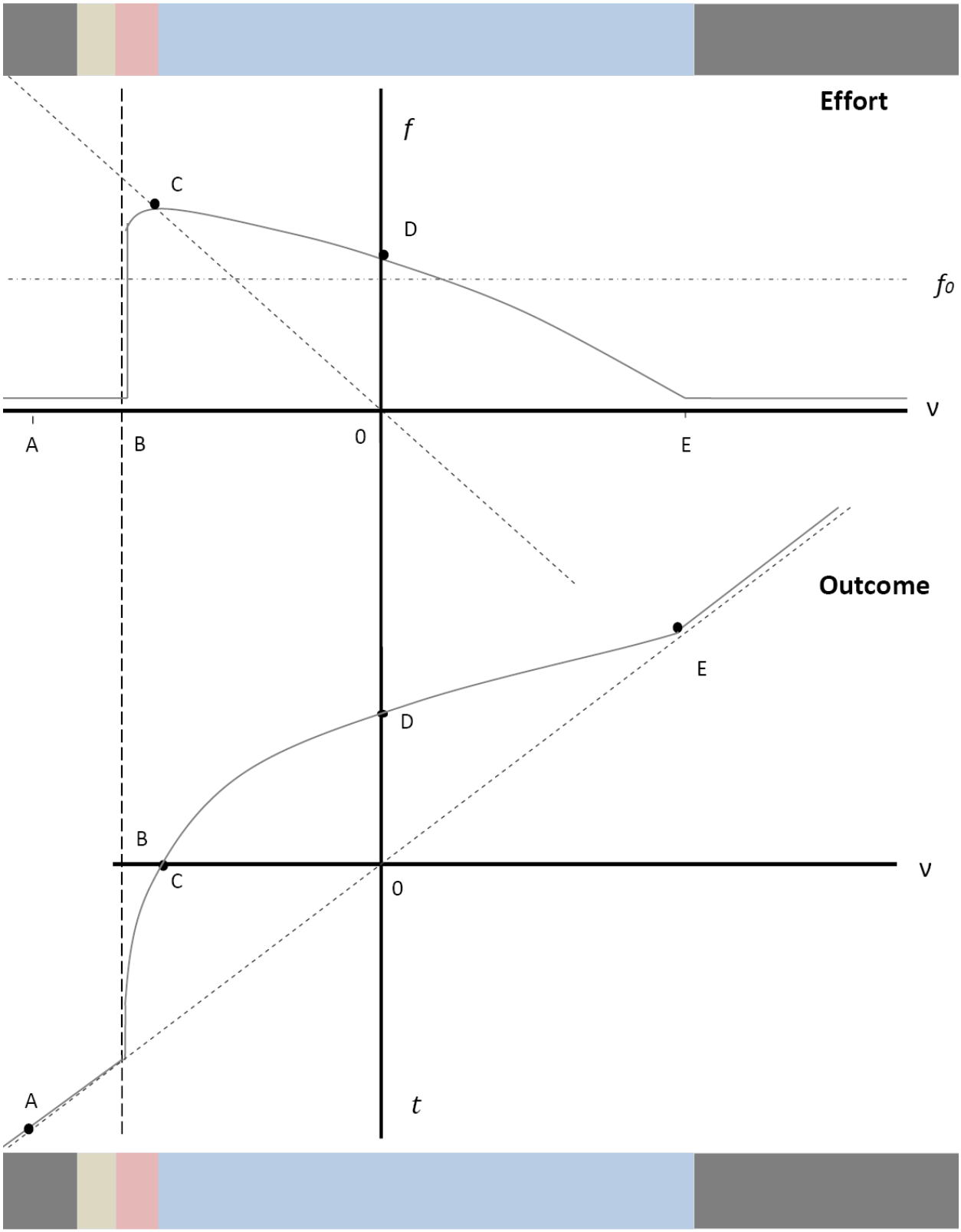
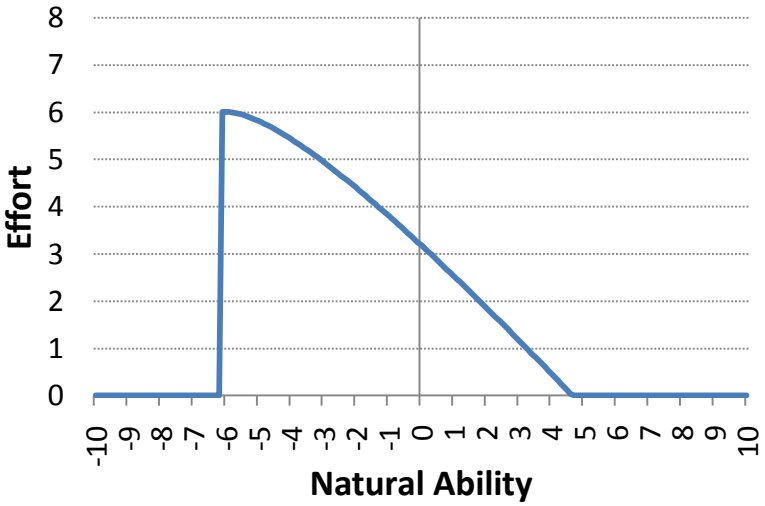
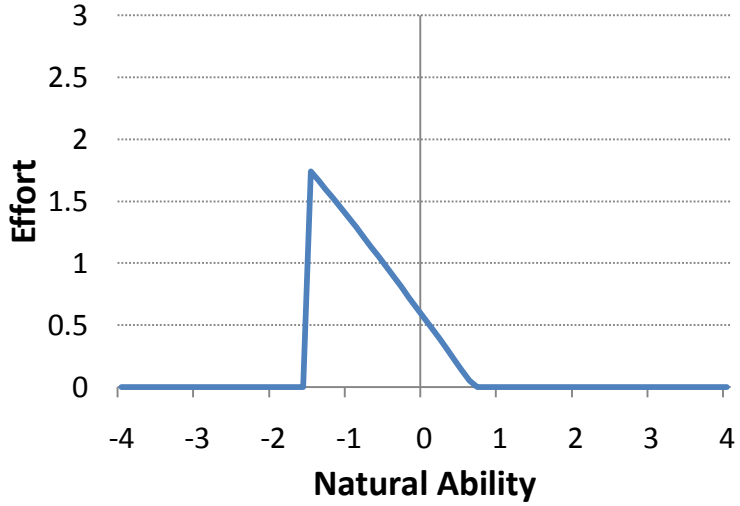


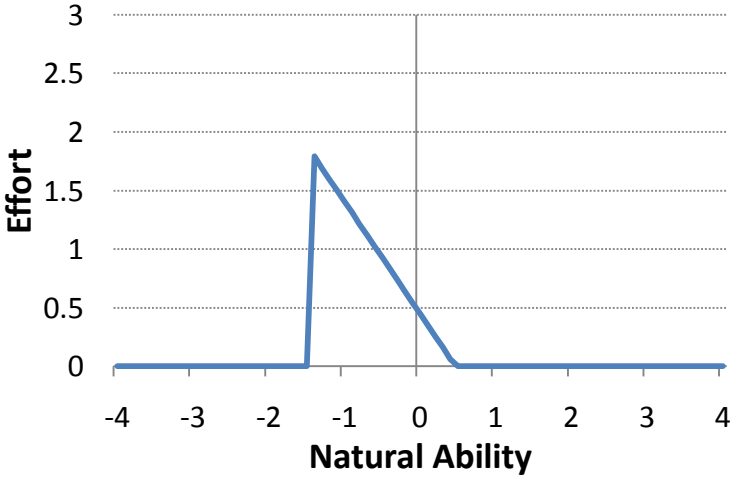
Figure 3: Theoretical Relation between Natural Ability and Effort under a Threshold Placed at Zero (ν ~ N(0,3)).



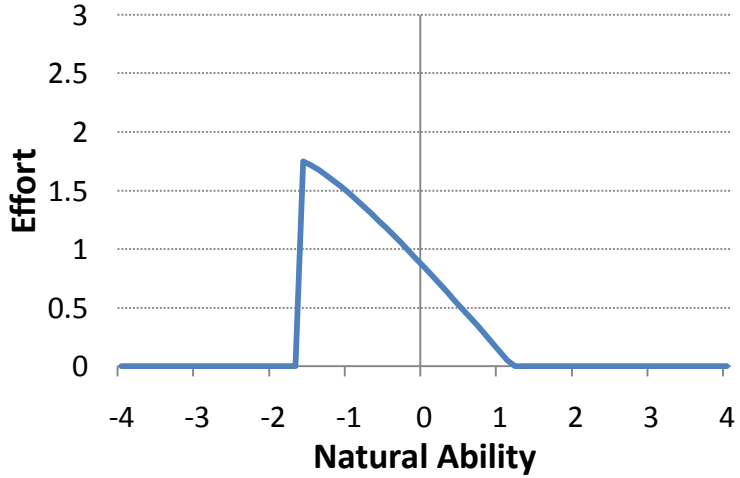
A. Parameter Values: $\gamma = 0.2$, $\sigma_\varepsilon = 3$, **P** = 5, mean $\nu = -3$.
$f_{\text{PERFECT}} = 6.34$, $f_{\text{IMPERFECT}} = 2.70$, $f_{\text{THRESHOLD}} = 3.73$.



B. Parameter Values: $\gamma = 1.2$, $\sigma_\varepsilon = 0.3$, **P** = 10, mean $\nu = -1$.
$f_{\text{PERFECT}} = 0.39$, $f_{\text{IMPERFECT}} = 0.38$, $f_{\text{THRESHOLD}} = 0.26$.



C. Parameter Values: $\gamma = 0.2$, $\sigma_\varepsilon = 0.3$, **P** = ½, mean $\nu = -1$.
$f_{\text{PERFECT}} = 0$, $f_{\text{IMPERFECT}} = 0$, $f_{\text{THRESHOLD}} = 0.23$.



D. Parameter Values: $\gamma = 0.6$, $\sigma_\varepsilon = 0.8$, **P** = 3½, mean $\nu = -1$.
$f_{\text{PERFECT}} = 0.16$, $f_{\text{IMPERFECT}} = 0.04$, $f_{\text{THRESHOLD}} = 0.35$.

Figure 4. Visual Overview of the Western States 100.

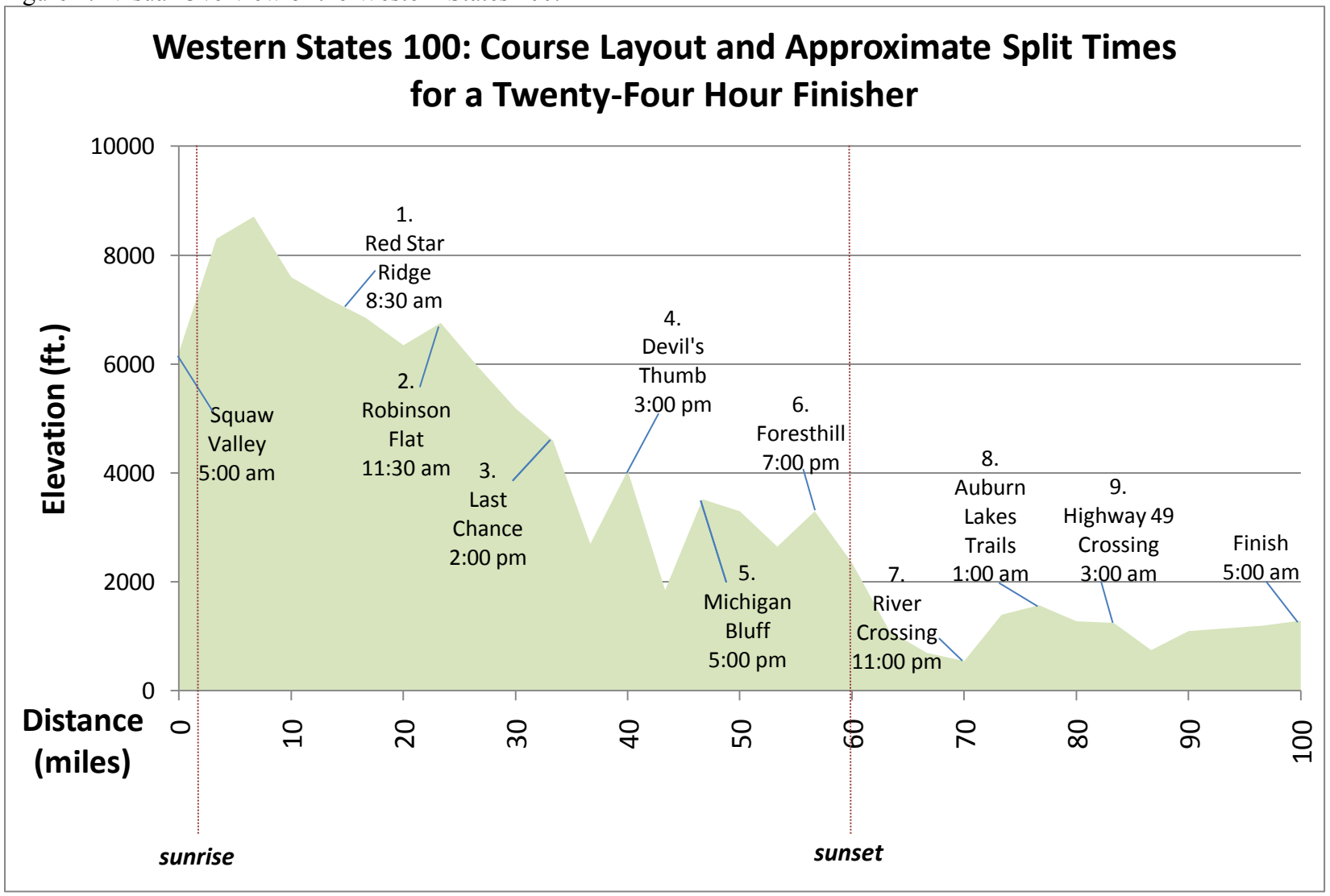
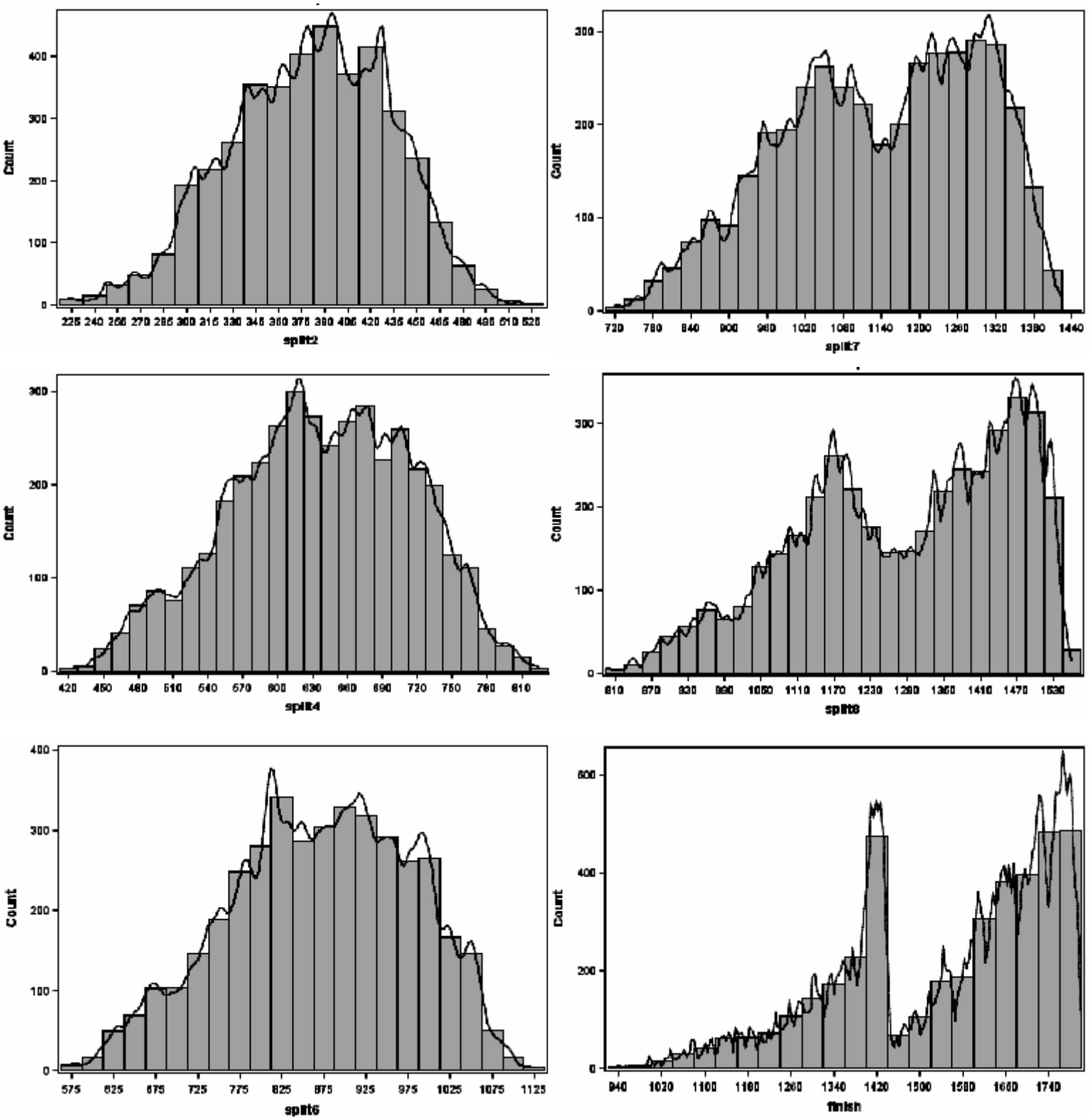


Figure 5. Histograms and Kernel Densities of Five Split Times and Finish Times (in minutes).



Left column: Times at splits 2, 4, and 6. Right column: Times at splits 7 and 8, and finish times.
A finish time of twenty-four hours corresponds to 1,440 minutes.

Figure 6. Split 6 and Finish Time Scatterplot, with Smoothed Mean (top) and Quantiles (bottom).

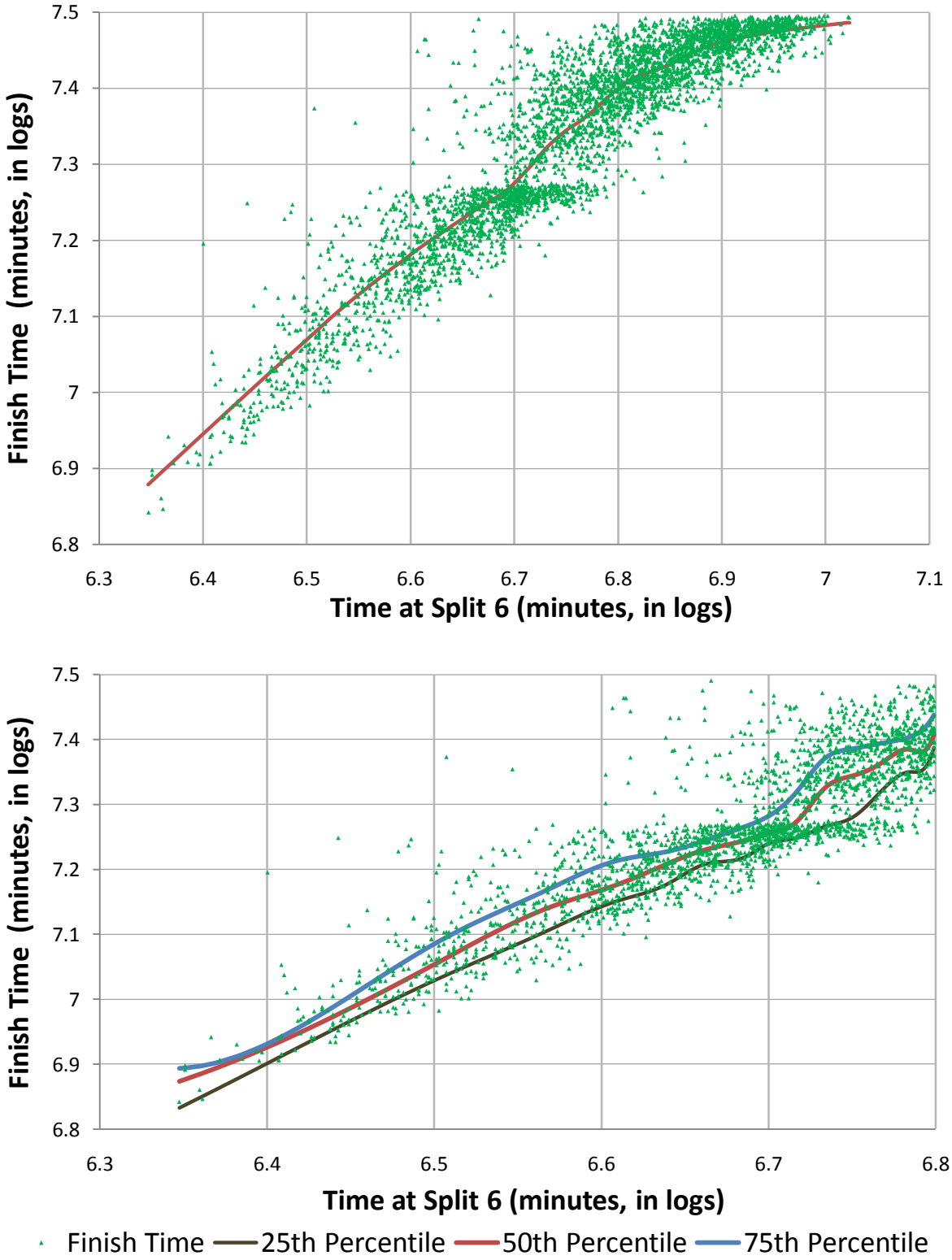


Figure 7. Smoothed Deviation of Actual Finish Time from Trend (95% Confidence Interval).

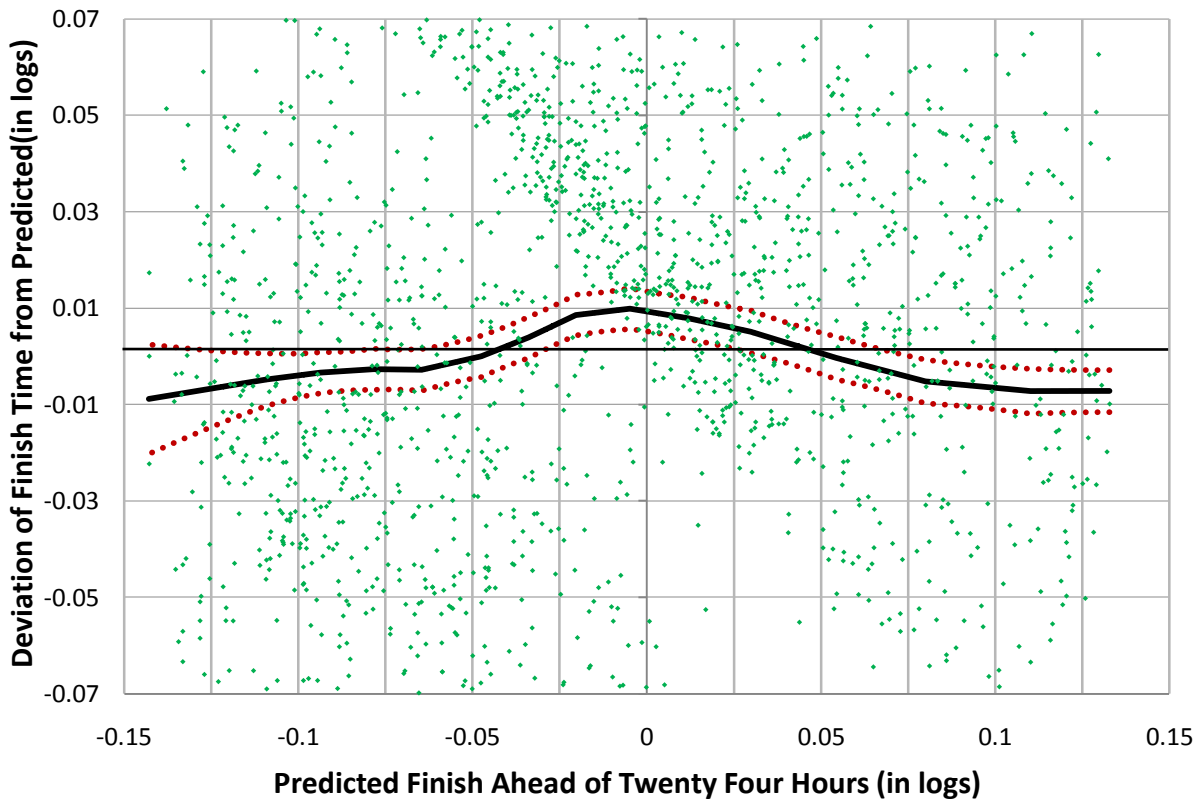


Figure 8. Effort Perturbation Surrounding the Threshold (95% Confidence Interval).

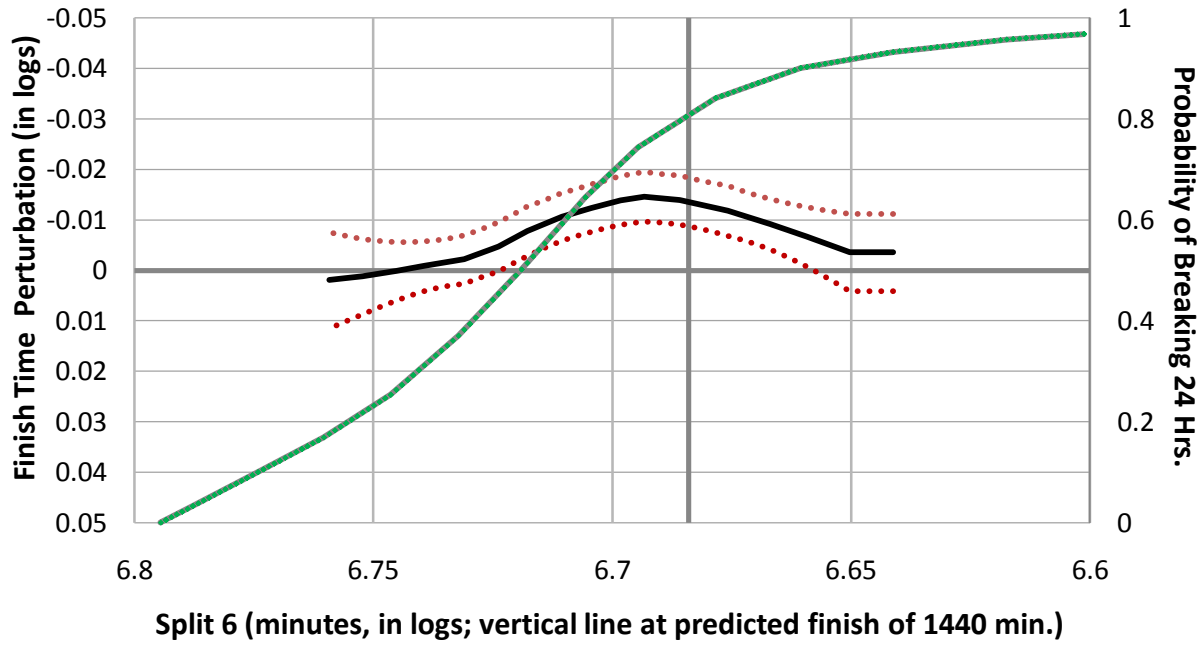
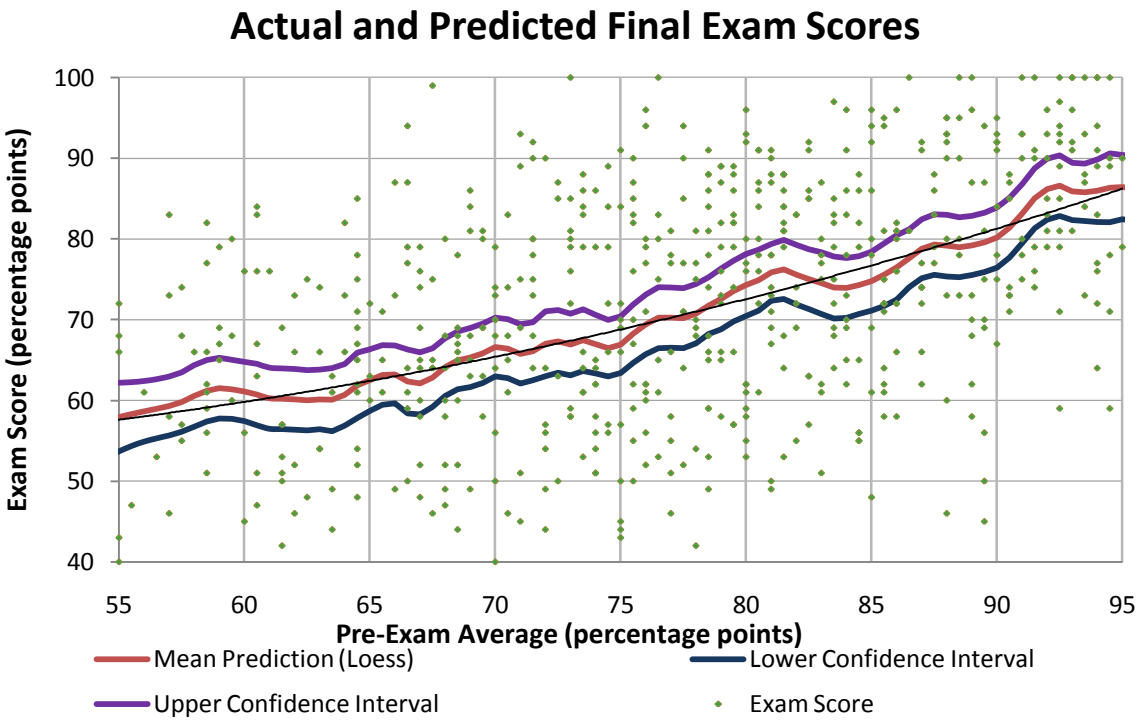


Figure 9. Abbreviated Results Portfolio: Grant (adapted from Grant and Green, 2010).



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	125 (0.19)	395 (0.60)	135 (0.21)	655 (1.00)

Results are for multiple sections of the author's Principles of Microeconomics class.

Figure 10. Reback (2008), Figure 2, and McEwan and Saltibañez (2005), Figure 4.

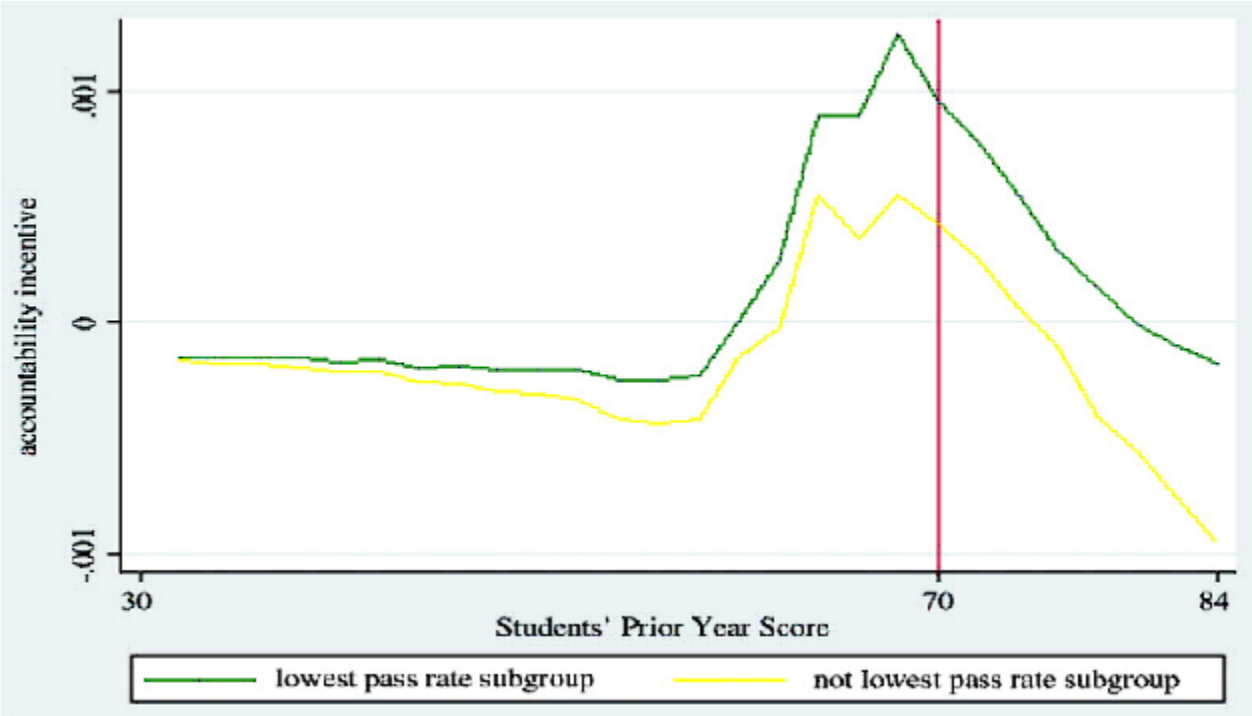


Figure 4
Teachers' initial points and classroom test scores (fitted values)