THE ESSENTIAL ECONOMICS OF THRESHOLD-BASED INCENTIVES[*]

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 when the agent crosses some pre-specified threshold. This paper comprehensively analyzes the effects of these incentive systems on effort, the net benefits of effort, and the accuracy of performance information that is provided to the public, and lays out methods for estimating each. These methods are then used to reveal the motivations and racing strategy of ultramarathoners trying to complete a one hundred mile race in under twenty-four hours.

JEL Codes: L83, C14, D10

Keywords: thresholds; behavioral incentives; ultramarathons

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

[*] This paper completes a trilogy on the economics of threshold-based incentives. In the application here, the threshold has strong incentive effects that conform to theory. In the application in the companion paper, Grant and Green (2013), there are no incentive effects. A related paper, Grant (2010), sketches out thresholds' effects on the distribution of performance when there is perfect measurement and population heterogeneity in the structural parameters. I am grateful to Mitchell Graff, Sohna Jaye, Lilly Park, Wade Pate, and Kevin Southerland for helpful research assistance, to participants at presentations at the North American Association of Sports Economists and Texas Camp Econometrics for helpful comments, and to Curtis Barton of the Seven Hills Running Club for guiding me to the application presented herein.

Rewards linked to the passing of a pre-determined threshold are a prevalent feature of economic life. Table 1 gives several examples. In business, thresholds separate workers who qualify for a performance bonus from those who do not. In the law, they classify certain offenses, such as drug possession or theft, into misdemeanors and felonies. In education they distinguish acceptable from unacceptable performance for students, schools, and school districts. There are thresholds for the collapse of ecosystems and for statistically significant research results. And research in behavioral economics, accounting, and psychology establishes that firms and individuals treat certain numerical values of performance, such as round numbers, as "focal points" that they then strive to meet.

This simple change from a standard, continuous reward structure dramatically affects its incentive properties. When the link between effort and reward is certain, the marginal benefit of improved performance is nil unless one crosses the threshold. When it is uncertain–as is typical in Table 1–expected marginal benefits rise and then fall in the neighborhood of the threshold. In both cases, incentive effects vary nonmonotonically and discontinuously with proximity to the threshold.

To date, however, development of the behavioral and normative properties of thresholds has been limited, with several papers pointing out that individuals who would otherwise fall "just short" will try harder in order to pass, and several others pointing out the potential perverse effects of incentives that do not reward improved performance once the threshold is reached. (Both types are well-represented in Table 1.) A full characterization of thresholds' incentive effects and normative properties is absent, as is a full development of the empirical methods used to estimate each.[1] The purpose of this paper is to remedy these gaps in the literature and provide a comprehensive theoretical

[1] In contrast to thresholds' well-studied cousin, regression discontinuity. Regression discontinuity designs measure the ex-post effect of an intervention by comparing outcomes on either side of an institutionally-imposed threshold separating those receiving treatment from those going without. Here, instead, the threshold is an incentive mechanism; the resulting discontinuity in effort, and its location, arise through optimizing behavior.

and empirical examination of threshold incentive systems.

Section I focuses on their behavioral effects. The theme of this section is simplicity: relatively little structure is needed to lay out essential theoretical predictions, test for the presence of threshold incentives, or estimate their effects on behavior involves. Compared to existing methods, this approach is more natural, more robust, and more rigorous.[2]

Section II focuses on their normative effects and estimation of the structural parameters underlying these effects. The theme of this section is the importance of uncertainty. Without it, thresholds cannot have beneficial normative properties and structural parameters cannot be identified; with it, both are possible.

Section III applies the methods from the first two sections to one of the most dramatic effort provision problems found anywhere, in ultramarathoning, where runners try to complete a one hundred race in less than twenty-four hours. In this application both the incentive effects and structural parameters are of interest, and our methods reveal non-obvious insights about human motivation and behavior. Section IV concludes.

We hope this paper will help the profession realize the full potential of this class of incentives, in both senses of the word. Thresholds are unlike typical labor supply problems in several ways: they yield a variety of unique behavioral predictions, can have beneficial normative properties that include,

---

[2] In some cases, the theoretical predictions developed in this section can be compared to existing estimates, sometimes successfully (Neal and Schanzenbach, 2010), sometimes not (McEwan and Saltibañez, 2005, which violates the "Peak Proximity Property" below). In other cases such a comparison is precluded by the estimation approach utilized, such as in Oettinger's (2002) study of grade incentives, which tests for incentive effects with a few dummy variables, or in Reback's (2008) study of accountability standards on school behavior, whose key independent variable (the "accountability incentive") imposes several of the properties tested here instead.

but go beyond, economic efficiency, and can reveal a large segment of the effort supply function, not just a local elasticity, without requiring instruments or temporal variation in prices. The existence and effects of thresholds in many areas of economic life are phenomena that deserve further exploration.[3]

I. Incentive Effects: Theory and Estimation.

Theory. An evaluator assesses a continuous behavioral outcome of interest. Under routine circumstances, the agent performs at some level of "ability," which represents a pre-existing combination of vigor, preparation, and natural endowment. If so motivated, however, the agent can give additional "effort" that improves performance on the outcome of interest. In many of the education scenarios discussed above, for example, ability would represent knowledge acquired through prior schooling and study, while effort represents additional tutoring or studying motivated by an upcoming exam or standardized test. In sales, ability could represent general selling skills and effort additional year-end selling diligence in order to satisfy an annual quota or achieve a bonus. In accounting, ability represents unmanaged earnings and effort investment in earnings management.

We allow the outcome as assessed by the evaluator to differ from the outcome anticipated by the agent at the time of evaluation. This could occur for many reasons, spanning most of the applications in Table 1. One is measurement or sampling error, as when a test asks only a subset of

[3] In a complementary paper, Dubey and Geanakoplos (2010) show that, in games of status, in which only one's relative rank matters, and binary effort, a threshold evaluation system can yield greater aggregate effort than a continuous system, and grading on a curve is never superior. In the model here, in contrast, absolute, not relative, performance matters and effort can take any nonnegative value. Rather than being placed in a separate literature review, other papers pertaining to the problem studied here are cited at appropriate points in the text.

all questions that might be asked. Another is error in self-regulation, as when a driver incorrectly self-assesses his blood alcohol level before driving. There could be some ambiguity in the assessment criteria. Or the outcome could be impacted by random factors beyond the agent's perception or control, as when unforeseen economic conditions cause the cancellation of some orders that had been placed for a salesman's product, or when an auditor unexpectedly requires an adjustment to reported earnings (Yim, 2013). This uncertainty in the link between effort and measured outcome is, we argue, of sufficient importance that it should be accounted for.

We intend to model the typical situation in which threshold incentives inspire, at most, perturbations in performance that are small relative to the variance in outcomes and which are confined to a small, local range of ability. This militates for simple functional forms that can be thought of, if desired, as first order approximations to more general alternatives. Accordingly, let the anticipated outcome, y, be the sum of ability, ν, and effort, f, and let the observed outcome be $Y = y + \epsilon$, where ϵ is independent, normally distributed error with a mean of zero and a standard deviation of σ_ϵ. The evaluator observes Y, but only reveals whether Y exceeds a pre-determined passing threshold that is normalized, for simplicity, to 0. Conditional on effort, the probability of passing is $\Phi((\nu+f)/\sigma_\epsilon)$, where Φ is the standard normal distribution function.

The reward for passing the threshold, **P**, can be set by a principal for the agent or determined by market forces and taken by each agent as given. The expected marginal returns to effort are then $\mathbf{P}\Phi'(\bullet)/\sigma_\epsilon$: a standard bell curve. A risk-neutral agent will equate these returns to the marginal costs of effort. The solution is easily determined when the costs of effort are specified as $C(f) = \kappa\bullet(\exp(\gamma f) - 1)$, where $\gamma>0$ represents diminishing returns or fatigue in the provision of effort and κ is henceforth normalized to one. The logged marginal expected returns to effort, $\ln(\mathbf{P}\Phi'(\bullet)/\sigma_\epsilon)$, form a quadratic

in f, while the log of marginal costs, $\ln(\gamma \bullet \exp(\gamma f))$, are a line.

For those agents who provide effort, the loci of points relating ability to effort forms a segment of a parabola in the $\{\nu,f\}$ plane that opens to the southeast (see the Appendix):

$$\begin{aligned} f(\nu;\gamma,\sigma_\epsilon,P) &= -(\gamma\sigma_\epsilon^2+\nu)+\sigma_\epsilon\sqrt{\gamma^2\sigma_\epsilon^2+2\gamma\nu+2\ln(0.4P/\gamma\sigma_\epsilon)} \\ &= -(\gamma\sigma_\epsilon^2+\nu)+\sqrt{(\gamma\sigma_\epsilon^2)[\gamma\sigma_\epsilon^2+2\nu+2\ln(0.4P/\gamma\sigma_\epsilon)/\gamma]} \end{aligned} \qquad \textbf{(1)}$$

A functional relationship like this is needed for structural estimation, discussed below, but not to lay out the basic properties of the threshold incentive effect. For this a few heuristics will do. By separating the essential intuitions from the functional form of the model, these generalize the model's predictions and reduce the assumptions required for estimation.

The heuristics can be articulated by depicting the derivation of equation **(1)** graphically. 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 (see also Becker and Rosen, 1992). This is reversed at the extensive margin, where it becomes 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 $y < 0$, 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 $y > 0$ at the extensive margin, the point of maximum effort occurs there, and effort declines thereafter.

Figure 2 depicts the resulting $\{\nu,f\}$ and $\{\nu,y\}$ loci for the non-trivial situation in which some agents exert effort. The relation between ability and effort is adequately described by five properties, depicted in the figure and described below, with proofs found in the Appendix.

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 has been previously shown by Oettinger (2002) and others.

2. **Sawtooth Property:** *Effort rises more quickly than it falls; that is, line BC in Figure 2a rises faster than line CE falls, so that the $\{\nu,f\}$ locus takes a sawtooth shape.* 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 ν 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 2a has a slope ≤ -1, so that those individuals who try the hardest–whose ability is argmax $f(\nu)$–have at least a 50% chance of passing the threshold.* This is a natural consequence of increasing returns to effort for $y < 0$.

4. **Precautionary Effort Property:** *Effort is positive at $\nu=0$.* Error in assessing y motivates precautionary effort to increase the individual's chances of passing (as for agent D).

5. **Stair Step Property:** *More able individuals have better outcomes than less able individuals; that is, $\Delta f/\Delta\nu > -1$ and $\Delta y/\Delta\nu > 0$.* Beyond point C, more able individuals work less and still have better outcomes. The $\{\nu,y\}$ locus always slopes upward, fastest near the extensive margin, like the sloping stair step in Figure 2b.

Each heuristic clearly extends to (and, to some degree, beyond) other functional forms satisfying the geometry of Figure 1: log-concave error and increasing marginal costs, both of which are commonly assumed in economics (Baghestanian and Popov, 2014). This is as good as one can do. Given the nature of the problem, the sweeping comparative statics of traditional price theory are not possible.

Estimation. These heuristics, and the local nature of the threshold incentive effect, are naturally suited to flexible, non-restrictive nonparametric or semiparametric estimation methods.

Density Estimation. The simplest and most intuitive test for the presence of threshold incentive effects relies on the expected bunching of agents just above the threshold, which should generate a discontinuity in the density of y.

The simplest way to test this intuition applies the "caliper method" (explicated in Gerber and Malhotra, 2008, and implemented in economics by Borghesi, 2008, and others) to the distribution of Y. In the absence of threshold incentive effects, "the conditional probability of observing an outcome that falls in a subset in a [suitably small] interval equals the relative proportion of the subset to the interval" (Gerber and Malhotra, 2008, p. 12). Thus, for an interval centered at the threshold, under the null hypothesis the population fraction of observations occurring above the threshold equals one-half. A one-sided rejection of this null implies threshold incentive effects are present. This test can be extended to ν-to-Y transitions or to more complex, more powerful nonparametric methods of estimating the density on each side of the threshold (McCrary, 2008).

Such methods have two problems, however, when the agent cannot perfectly predict whether he will pass ($\sigma_\epsilon > 0$). First, as Figure 2 demonstrates, the discontinuity in y almost never occurs at the threshold, but at a value that depends on unknown structural parameters and thus cannot be pre-specified (and which can, in fact, be negative, as in the figure). Second, the density of Y, unlike that of y, is not (in general) discontinuous, because of the presence of uncertainty.[4]

[4] This is easily shown for the normal errors assumed here. McCrary was careful to emphasize his test's validity held only in the case of certainty–"perfect manipulation" of the running variable, in his parlance–but this warning has not always been heeded in practice. The act of manipulating the running variable in a regression discontinuity framework is itself subject to threshold incentive effects, with the un-manipulated variable serving as ability and manipulation serving as effort.

Thus, while density tests are often useful as a "first cut" at the data, they should be supplemented when possible with an analysis that directly relates ν to Y. This not only yields a natural test for the presence of threshold incentive effects that avoids these problems, but also reveals the incentive effect itself.[5]

Least Squares Estimation. A semiparametric regression that estimates the $\{\nu, f\}$ or $\{\nu, y\}$ loci directly, generating empirical results in the format of Figure 2, is easily implemented and allows the properties above to be tested, formally or informally, rather than imposed.

Consider a parametric regression that assumes a threshold has no incentive effects. If so, outcomes should be a smooth function of ability (in the colloquial, not mathematical, sense)–that is, a trend, such as the linear relationship assumed in our theoretical model. Allowing Y and ν to be measured in different units, and including control variables X and error ξ, this linear relationship is:

$$Y = \alpha + \beta\nu + \lambda X + \xi \qquad \textbf{(2)}$$

We can treat the adequacy of this specification as the null hypothesis in a specification test for the presence of threshold incentive effects. The alternative is that this parametric relation is inadequate, because effort is systematically related to proximity to the threshold. If so, the residuals near the threshold should be "autocorrelated." Absent controls, a simple test is based on A, the average squared error in equation **(2)**, and B, one half of the mean squared difference between

[5] Recent working papers by Yim (2013) and Allen et al. (2014) model the bunching implied by threshold incentives based only on the observed density. This approach's limitations stem from the absence of an observed ability measure: estimation is replaced with simulation, in which additional free parameters are calibrated in order to try to reproduce the distribution of observed outcomes.

adjacent values of Y, after being placed in ν-order:

$$Z = \sqrt{S}\cdot(A - B)/B \sim N(0,1) \quad \textbf{(3)}$$

where S is the number of observations (Yatchew, 1998). The null is rejected for sufficiently large values of the test statistic Z. When controls are present, practical alternatives are provided by Henderson and Parmeter (2014), Pagan and Ullah (1999), and Yatchew (1998).

This test can be strengthened using a priori information on the ability domain, $\nu_L \leq \nu \leq \nu_H$, over which threshold incentive effects may be expected to appear.[6] The null should be rejected for this domain only. On its complement, equation **(2)** should suffice. If the appropriate null is rejected, the effort perturbation $g(\nu)$ is then estimated semiparametrically, as follows:

$$Y = \alpha + \beta\nu + g(\nu)\cdot 1(\nu_L \leq \nu \leq \nu_H) + \lambda X + \xi \quad \textbf{(4)}$$

A discontinuity in performance is not formally specified; it should reveal itself as a sharp rise in performance for some value of ν, ν_{EX}, where $\nu_L \leq \nu_{EX} < 0$. We suspect this will usually be adequate. If not, one can use the structural or quasi-structural models introduced below.

This equation is easily adapted to a single-index model, when ν cannot be directly observed but can be predicted from other observed variables, or to a discrete choice framework that only needs data on $1(Y \geq 0)$.[7]

---

[6] Ideally, this domain is chosen without reference to the data. But, as emphasized by Hardle and Horowitz (1994) in a related application, sometimes this can be difficult to do. This issue is addressed in the empirical application below by choosing an interval that is a "round number" (0.1 log points) that is centered on a ν-value that is itself a "round number" (a multiple of 0.1 log points).

[7] The single index model is $Y = \nu + g(\nu)\cdot 1(\nu_L \leq \nu \leq \nu_H) + \lambda X + \xi$, with $\nu = \alpha + \theta Z + \zeta$, where Z are observed predictors of ability, ζ is an error term, and α and θ are coefficients. But two caveats

Quantile Estimation. It is important to recognize that equation **(4)** distinguishes between the performance-predicting information that is, and is not, available to the agent. The index ν captures everything known to the agent ex ante, with controls limited to factors not known to the agent prior to evaluation. In the companion paper, for example, semester dummies control for inter-semester variation in the difficulty level of each instructor's final exams, assuming students know only the average difficulty level of these exams, not the inter-semester deviations.

Furthermore, if ν is not perfectly observed (or perfectly predictable), the error term could reflect ability information that was known to the agent ex ante but not observed by the econometrician. If so, semiparametric mean regression is problematic. To see this, scale Y so ß=1, and decompose the error term in equation **(4)**, ξ, into components reflecting privately-known ability, π, and true random error, ω. Both are mean zero and uncorrelated with ν. Then:

$$\begin{aligned} Y &= \alpha + \nu + \pi + g(\nu + \pi) + \lambda X + \omega \\ E_{\pi}(Y|\nu,X) &= \alpha + \nu + E_{\pi}[g(\nu + \pi)] + \lambda X \end{aligned} \qquad \textbf{(5)}$$

Semiparametric mean regression estimates the perturbation in the second line. This does not equal g, but rather a convolution of g with the density of π, which smooths out, or disperses, the original function, so that it is diminished on the vertical scale and overly broad on the horizontal. (Fortunately, if π is normally distributed, at least, the five properties articulated above still apply.)

---

in extending equation **(4)** to a probit or logit model should be noted. First, the quantile regression that is advocated below cannot be estimated. Second, estimates of the $\{\nu, Y\}$ trend, ß, in the underlying latent variable can be far less robust, as almost all low-ν observations fail, while almost all high-ν observations pass. Estimates of the effort perturbation near the threshold are affected accordingly. Both limitations pertained in probits estimated on the ultramarathoning data below, which predicted whether a contestant completed the run before the course closed, or whether a finisher broke the twenty-four hour threshold. These latter results resemble those in Figure 6b.

This problem can be addressed with semiparametric quantile regression, which partly accounts for private information about ability. To see this, let ω vanish and π be homoskedastic in ν. Given ν, one's performance ranking is a monotonic function of π. Quantile regression thus conditions on π, so g is recovered nonetheless. Even when ω is nonzero, if private information is substantial, quantile regression should still be quite helpful. In addition, if only some agents are motivated by threshold incentives, incentive effects may be revealed at high quantiles even when they are absent elsewhere. As there is little cost to using quantile regression, we recommend doing so routinely.[8]

II. Normative Effects: Theory and Estimation.

Thresholds' normative effects are best examined using a strict interpretation of the model above, in which ν represents pre-existing "natural ability," the market value of a unit of y is $\mathbf{p}$, and ϵ represents measurement or sampling error in evaluation. (Some conclusions that follow extend to other interpretations.) We compare a threshold to direct, continuous reporting of performance. Under the threshold the market sets the value of passing at $\mathbf{P} = (\bar{y}_{PASSERS} - \bar{y}_{NONPASSERS})\mathbf{p} = \Delta\bar{y}\mathbf{p}$.

If the evaluator perfectly measures performance ($\sigma_\epsilon = 0$) and reports it directly, each agent's effort maximizes the difference between its rewards, $\mathbf{p}f$, and its cost, $C(f)$. If $\mathbf{p}$ reflects the marginal social benefit of y, then continuous, direct measurement of y provides ideal information to the market and appropriate incentives to the agent, and thresholds are unnecessary (see Costrell, 1994).

---

[8]The one threshold study using both least-squares and quantile methods (Oettinger, 2002) supports this recommendation: estimates from the former were insignificant, but not those from the latter. Nonparametric quantile estimators now can be found in Limdep, while Hayfield and Racine (2011) present a kernel-based package, np, for the programming language R; the spline-based methods used here (for example, Wang and Yang, 2009) can be implemented in SAS.

When output is measured with error, on the other hand, direct measurement exhibits the classic signal extraction problem: variation in the measured outcome is attributable partly to population variation in *y* and partly to error. Assume henceforth that ν is also normally distributed (throughout the population) with standard deviation σ_ν. The market price of a unit of *Y* is then $\mathbf{p}\sigma_\nu^2/(\sigma_\nu^2+\sigma_\epsilon^2)$;[9] the price of a unit of effort is attenuated by $\sigma_\nu^2/(\sigma_\nu^2+\sigma_\epsilon^2) < 1$. Effort diminishes accordingly. It is possible to increase the accuracy of the information provided to consumers, the total effort elicited by agents, or the net benefits of effort (efficiency). Under the right conditions thresholds can do one or all of these. *Thresholds can be justified by imperfect information.*

Theory. We now sketch out the conditions under which this occurs.

Motivating. By leveraging the divergence in performance between passers and non-passers, $\bar{y}_{\text{PASSERS}} - \bar{y}_{\text{NONPASSERS}} \equiv \Delta\bar{y}$, thresholds can magnify the returns to effort, thus increasing aggregate effort. That is, the expected marginal returns to effort under a threshold, $\mathbf{P}\Phi'/\sigma_\epsilon \equiv \mathbf{p}\Delta\bar{y}\Phi'/\sigma_\epsilon$, can exceed those under direct measurement, $\mathbf{p}\sigma_\nu^2/(\sigma_\nu^2+\sigma_\epsilon^2)$, for most agents. If these returns are not magnified too greatly, so that effort is overprovided, efficiency also increases.

These motivational effects can be characterized in terms of the "potency" of the incentive, $\mathbf{p}/\gamma$. Incentives are more potent when they are stronger (higher **p**) or when agents respond more to them (lower γ). For any realistic value of $\sigma_\epsilon/\sigma_\nu$,[10] the Appendix derives a range of potency for which

[9] In the equilibrium supported by this price, each person's effort is optimal given everyone else's choices. As each person provides the same amount of effort, the variance of *y* ex post equals the variance of ν.

[10] These results require $1.2\sigma_\epsilon < \sigma_\nu$ (see the Appendix). This is realistic–error is generally much less variable than ability. Alternatively, when error is large, effort under direct measurement falls far short of efficiency, but a threshold may also be impotent: passing is primarily due to luck, not

thresholds alone induce effort, and a subset of that range for which that effort is guaranteed to be inefficient. Simulations that build on these findings generate "bands of potency" under which mean effort under a threshold is efficiently, and inefficiently, greater than that under direct measurement. Figure 3 depicts both for a typical case, in which $\sigma_\epsilon/\sigma_\nu = ½$. Each band slopes upward at an angle of roughly 45°, along which potency is constant. For other cases the results are qualitatively similar, with the set of information- and efficiency-improving parameters shrinking as $\sigma_\epsilon/\sigma_\nu$ falls.

Figure 3 shows that thresholds' motivational properties are strongest when potency is sufficiently mild, increasing both effort and efficiency. At higher levels of potency, moving to the upper left in the figure, thresholds continue to increase mean effort, but efficiency falls. Thresholds can be a blunt instrument, underincentivizing some agents while overincentivizing others. At still higher levels of potency, thresholds decrease mean effort and efficiency. At this point most agents are essentially infra-marginal, resigned to failing or clustered at large values of y, where they are quite likely to pass. In consequence, increases in potency do not call forth much additional effort, and direct measurement, which does not have this problem, becomes superior.

Informing. Information may be desired about ability, ν, as in a signaling model, or about true performance, y. Thresholds can help with the latter but not the former.

Thresholds are problematic for signaling because the effort of passers is negatively related to ability. Low-ν individuals exert great effort to pass, while high-ν individuals exert, at most, a little precautionary effort. This negative relation widens the variation in ability conditional on passing. This is avoided, along with some truncation error, by using direct measurement.

---

performance, so $\Delta\bar{y}$ can become small. There is then no systematic method of constructing examples of effort-increasing thresholds, but when they do, efficiency usually increases as well, as in Figure 4a.

For inferring performance, however, the opposite is true: the negative relation between passers' effort and ability diminishes the variance of y conditional on passing. 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 y for passers, var($y|Y>0$), is less than var($y|Y$) when performance is measured directly.

As Figure 3 shows, such outcomes are easily generated when measurement is noisy and incentives are potent. The true performance of those agents who exert effort is:

$$y(\nu;\gamma,\sigma_\epsilon,\boldsymbol{P}) = \nu + f = -\gamma\sigma_\epsilon^2 + \sigma_\epsilon\sqrt{\gamma^2\sigma_\epsilon^2 + 2\gamma\nu + 2\ln(0.4\boldsymbol{P}/\gamma\sigma_\epsilon)} \qquad \textbf{(6)}$$

For sufficiently high rewards or low fatigue, $\gamma\nu << \ln(0.4\mathbf{P}/\gamma\sigma_\epsilon)$, so these agents' performance is weakly related to ν. If there aren't too many inframarginal passers, the spread in true performance conditional on passing will be small.

The conditions under which thresholds improve information accuracy are clearly distinct from those under which effort and efficiency improve. Yet, as Figure 3 shows, a threshold can be both motivationally and informationally superior.

Identification and Estimation. Estimates of the structural parameters are needed to quantify these normative effects, or can be useful in themselves. Under the right circumstances, the data identify $\{\mathbf{P}, \gamma, \sigma_\epsilon\}$; given the distribution of Y and the estimate of σ_ϵ, $\Delta\bar{y}$ and hence $\mathbf{p}$ can be calculated. Note that neither $\mathbf{P}$ nor the cost function $C(f)$ is given in dollar terms, as this function's constant, κ, has been normalized to one. Then one cannot determine the magnitude, only the sign, of the effect of the

threshold on efficiency. But if the nominal reward for passing is observed, then κ equals the ratio of this to the structural estimate of **P**. Then C(*f*) can be put in dollar terms and the dollar magnitude of the effect on efficiency calculated.

Identification of the structural parameters can be strong, weak, or non-existent, depending on the circumstances. They are not identified from the points on the ability-effort profile alone. These depend on just two composite parameters, $k1 = \gamma\sigma_\epsilon^2$, $k2 = \ln(.4\mathbf{P}/\gamma\sigma_\epsilon)/\gamma$, in the second line of equation **(1)**. There are three cases, which can be categorized by the nature of ϵ.

Case 1: No Uncertainty. We have already discussed the normative properties of this case, but there still may be interest in the remaining structural parameters. Unfortunately, as $\sigma_\epsilon \to 0$ equation **(1)** devolves to $f = -\nu$, and provides no value in identifying these parameters. The location of the extensive margin $\{\nu_{EX}, f_{EX}\}$ provides some value, but not enough: $-\nu_{EX} = f_{EX} = \ln(\mathbf{P})/\gamma$. Neither **P** nor γ is identified. Thus we have a surprising result: *uncertainty is required in order to identify the structural parameters of threshold-based incentive problems.*

Case 2: No Private Ability Information. Here identification can be completed from the proportions of agents passing the threshold for various values of ν. Given true performance *y*, the probability of passing the threshold is $\Phi(y/\sigma_\epsilon)$. This suggests a probit model with a dummy for passing the threshold as the dependent variable and the estimate of *y* as the independent variable. If $y(\nu)$ is consistently estimated as outlined above, the inverse of the slope coefficient in this probit model consistently estimates σ_ϵ.[11] Structural estimation is not required to estimate the remaining parameters; the method of moments applied to the semiparametric ability-effort profile will do.

[11] Surprisingly, the effect of the threshold on mean effort can then be signed given the distribution of *Y*, without knowing γ or **P** (see the Appendix).

Case 3: Private Ability Information. Here structural estimation of the full ability-effort profile is required; identification is completed via the location of the extensive margin. Here, however, identification can be weak. As the Appendix shows, to the second order, the location of the extensive margin is also governed by the aforementioned composite parameters. Thus, disparate combinations of structural parameters associated with similar {*k1*, *k2*} values can generate similar ability-effort profiles (especially when incentives are not too potent–see the Appendix). Then structural parameter estimates will be imprecise and normative effects unclear. Panels b-d of Figure 4 illustrate this phenomenon, depicting nearly identical ability-effort profiles that are generated from varying sets of parameter values with divergent efficiency properties. Compared with direct measurement, threshold effort is inefficiently underprovided in panel b, efficiently provided in panel c, and inefficiently overprovided in panel d.

This problem shapes our approach to estimation: along with a structural model, we introduce a "quasi-structural" model that sacrifices precision in the specification for greater precision in the estimates. The structural econometric model inserts the first line of equation **(1)** into equation **(4)**:

$$Y = \alpha + \beta\nu + 1(\boldsymbol{P}\Phi((\nu + f(\cdot))/\sigma_\epsilon) > C(f(\cdot))) \cdot f(\nu; \gamma, \sigma_\epsilon, \boldsymbol{P}) + \lambda X + \xi \qquad \textbf{(7)}$$

where $f(\bullet) \geq 0$, $\Phi(\bullet)$, and $C(\bullet)$ are defined in Section I. The effort discontinuity at the extensive margin is governed by $1(\bullet)$. The quasi-structural model inserts the second line of equation **(1)** into equation **(4)**, instead, and appends a free parameter, ν_{EX}, for the location of the extensive margin:

$$Y = \alpha + \beta\nu + 1(\nu > \nu_{EX}) \cdot f(\nu; k1, k2) + \lambda X + \xi \qquad \textbf{(8)}$$

where *k1*, *k2* are defined above. Either model can be estimated in least squares or quantile regression.

III. Application to Ultramarathoning.

Our empirical application, to ultramarathoning, utilizes simple implementations of our semiparametric and structural models, while representing the major estimation issues discussed above.

Data and Institutional Details. California's venerated Western States 100 (WS100) admits roughly 370 runners each year, by lottery, from about one thousand applicants. Each applicant must qualify by demonstrating the capacity to complete the WS100–though not necessarily quickly. The one hundred mile course closes after thirty hours, and a coveted medal is presented to finishers under twenty-four hours. Only one runner in four meets this standard, so this medal is a mark of distinction. As we shall see, runners are highly motivated to meet this threshold.

With a few exceptions, such as years with wildfires, the WS100 has run the same course since its inception in 1977. Finish times and "split" times, taken at nine aid stations spread throughout the course, are recorded on the run's web site (www.ws100.com) for its 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 finishers over the period 1986-2006. Split times are recorded in minutes, finish times in minutes and seconds.

Figure 5 illustrates the course layout and the distribution of times at selected aid stations and at the finish, for the full sample, with a simple histogram and a more precise kernel density. (The scale of the horizontal axis, suppressed for clarity, increases as the race progresses.) The layout and the distributions both demonstrate that the course is effectively run in two stages. The first stage, through split six, features high elevations, steep gradients, and temperature extremes of mountain cold

and daytime heat. The split times here are almost normally distributed. The second stage, after split six, begins around nightfall, after which the course cools, drops in altitude, and flattens out. Here each runner is allowed a companion, or "pacer," because of their mental and physical deterioration at this point in the race. Immediately the distribution of times begins to bifurcate. At the finish it is almost divided in two, one bunched ahead of twenty-four hours, and another ahead of thirty hours, when the course closes. All this suggests contestants run the first stage of the race at a reasonably even pace, generating the ever-widening, normal split distributions, and then, with their pacers, tweak their times during the second stage to try to satisfy one of the two thresholds.

There is abundant precedent in long-distance running for a two-stage strategy of this sort. This is because of a feedback effect: running too fast too early induces physiological deterioration that hampers performance later on. This is most visible at "middle distances" such as five or ten kilometers. There, most of the race is run at a fairly even pace that relies on easily-sustained aerobic energy. Then, toward the end, the runner begins a "kick" that draws on a limited store of anaerobic energy. The waste thereby released, lactic acid, quickly accumulates in the muscles, diminishing performance, so the kick is vigorous but brief. A similar principle pertains in ultramarathoning, though far more gradually, and less because of anaerobic energy use than because of the general physical and mental deterioration that accrues over such long distances (Weir et al., 2006). Allen et al. (2014) confirm the use of a similar strategy in the 26.2-mile marathon, and argue that runners strive to beat "round number" finish times, such as twenty-four hours, because of prospect theory.

Our empirical model is well-suited to this two-stage strategy. The sixth split time, at the end of the first stage of the race, represents ability, ν (lower times are better). Output, Y, is the finish time, and threshold-motivated effort, $g(\nu)$, generates unexpectedly fast finish times for runners that

were just on or just behind schedule to meet the threshold. The fatigue parameter, γ, represents the increasing difficulty of further increasing one's pace, while the error term, ϵ, captures the major source of uncertainty: imprecision in regulating one's pace and predicting one's finish time mid-race.[12]

The use of the sixth split time as our ability proxy is also supported in preliminary regressions using the sample and specification defined below. Parametrically, it alone explains 78% of the variance in finish times; the first six splits together explain 80%. Furthermore, there is no sign of an effort perturbation in the sixth split times themselves. These findings suggest they are reasonable, though not perfect, indicators of participants' general fitness and endurance. Some of the remaining variance in finish times, however, may reflect private ability information: differences in the degree of fatigue each participant perceives at the end of the race's first stage. Two runners with identical sixth split times should not expect to finish together if one feels fresher than the other. These differences in perceived freshness could result incidentally, from feeling strong that day, or deliberately, from differences in racing strategy. Either way they would by revealed by, and necessitate the use of, quantile estimation of equation **(4)**, which would differ materially from estimates of the mean.

Estimating the Incentive Effect. The caliper test easily supports the importance of the twenty-four hour threshold. In Figure 5, a total of 97 runners finish at most ten minutes ahead of this threshold, while only 19 runners finish at most ten minutes behind it, a highly significant difference.

We thus relate finish times to the sixth split times, restricting the sample to those 2,273

[12] Year dummies, which capture small differences in mean finish times conditional on the time at the sixth split, are not included as controls, because their values are not wholly unknown to the runner: they are, instead, gradually revealed to the runner in the split times observed during the second stage of the race. Their inclusion does not alter the general shape of the estimated effort perturbation, however; and materially strengthens the specification tests below.

individuals whose sixth split is less than fifteen hours, for their finish times are very rarely censored (see Figure 6a). Both times are measured in logarithms, which best fit the data, and the linear relation is empirically supported by the insignificance of a quadratic term, when added. The $g(v)$ estimates can be interpreted as percentage changes in time or overall pace; suitably transformed, they also estimate the ratio of the second-stage pace to the first-stage pace, how much the runner "sped up".[13]

We begin with a parametric regression analysis, as in equation **(2)**, which yields an estimated intercept of -0.053 (standard error 0.08) and trend of 1.10 (standard error 0.01). The relevance of threshold incentive effects is supported by the specification test in equation **(3)**. The test statistic of 1.45 ($p = .07$) marginally rejects the null hypothesis of no misspecification. This regression predicts a twenty-four hour finish for a logged split time of 6.68, and the residuals suggest unusually strong finish times in this (and only this) neighborhood, associated with logged split times of roughly 6.65 to 6.75. Splitting the sample, and conducting the specification test separately on observations falling within and outside this domain, the null hypothesis is rejected for the former, with a test statistic of 1.70 ($p = .04$), but not the latter, with a test statistic of 0.47.

We structure the nonparametric term accordingly in the mean regression of equation **(4)**. Figure 6b presents estimates of the effort perturbation $g(v)$ from this regression, conducted with a loess smoother, with the bandwidth chosen by cross validation. Each axis has been placed in descending order, so outward movements on each indicate better performance or ability, as in Figure 2. Performance improves by as much as 1.5% for threshold-motivated runners. The Peak Effort Property and the Precautionary Effort Property are transparent, while the Peak Proximity Property

[13] One can easily show that a finish time perturbation of -1% is associated with a change in overall race pace of -1% and a change in the ratio of the second-stage pace to the first-stage pace of (total time/second-stage time)%, which, for the average finisher, is about -2.25%.

is also confirmed: at the point of maximum effort the runner has a three-fourths chance of passing the threshold.[14] The slope of the declivity is less than one in absolute value, supporting the Stair Step Property, and is one-third less than the slope of the acclivity, supporting the Sawtooth Property.

The sixth split time may be an imperfect indicator of ability, however, as noted above. If so, quantile regression of equation **(4)** should yield improved estimates of the effort perturbation $g(v)$. It does indeed. The trend, $\hat{\beta} \approx 1.15$, is virtually identical at the 25th, 50th, and 75th percentiles (implying ξ is homoskedastic). Thus all three perturbations can be depicted by a single graph of these smoothed, detrended quantiles (smoothed quantiles of $Y - \hat{\alpha} - \hat{\beta}v$), shown in Figure 7a. Our five heuristics are clearly satisfied. At the extensive margin, increases in effort shave 3% off the finish time, an increase in the second-stage pace of almost one minute per mile. The modest mean threshold effects in Figure 6b are, indeed, a convolution of much stronger effects dispersed across runners at different quantiles. The similarities in the size and shape of each effort perturbation suggest that ultramarathoners are more alike than different in their motivations and use of the two-stage strategy.

The sharp onset of each effort perturbation indicates the location of the extensive margin, which turns out to be nearly thirty minutes "off pace" for a twenty-four hour finish. Runners who are further behind than that at the sixth split do not try to meet the threshold, and the absence of a material "dip" in Figure 7a below the extensive margin indicates these runners' times do not suffer in despair. The remaining runners do try to meet the threshold. Most are successful, as the histogram in Figure 5 and scatterplot in Figure 6a make clear.

Figure 7b illuminates the strategy runners employ to achieve this end. This figure conjoins

[14] This finding implies that maximum effort occurs at the extensive margin, while the gentle upward slope of the perturbation suggests that it occurs at an interior solution instead, as in Figure 2. This seeming contradiction is reconciled in the quantile regression below.

two effort perturbations: the eighth split time on the sixth split time, in the right half of the figure, and the finish time on the eighth split time, in the left half, both estimated as in Figure 7a, at the 50th percentile. Both perturbations are sizeable, but the latter is higher and "sharper" than the former, indicating that threshold-motivated runners deploy increasing, increasingly focused effort throughout the second stage of the race. This affirms our theoretical model: the effective degree of uncertainty about one's eventual finish time decreases as one progresses through the race, and this diminished uncertainty yields higher, sharper ability-effort profiles.

The allure of the 100-miles-in-24-hours standard cannot be overstated. The web site RealEndurance.com (realendurance.com/list.php) presents finish time histograms for every ultramarathon in the U.S. Virtually all show bunching just ahead of twenty-four hours.[15] While the incentive effects documented here may be augmented by the receipt of a medal and the prestige of the WS100, the prevalence of this phenomenon across a wide variety of ultramarathons indicates that most motivation is intrinsic. The structural model below provides further insight into that motivation.

Structural Estimation. The normative issues related to market-based thresholds clearly do not apply to the WS100, where finish times are accurately measured and freely published, and output is not sold. Structural estimation is useful nonetheless, distinguishing the mental and physical primitives underlying the threshold incentive effect. Because of the prevalence of private information about v,

[15] One can also consider the dual: "twenty-four hour races" in which the objective is to run as far as possible. The most venerable of these is the Sri Chinmoy Self-Transcendence Ultra Classic, run around an indoor track in Ottawa, Canada, with results recorded in kilometers. A histogram of the last five years of finishes reveals a broad underlying normal distribution centered around 135 km, punctuated by a sizeable cluster of finishers in the right neighborhood of 100 km, and another in the right neighborhood of 161 km–that is, 100 miles.

we use quantile estimation of the 50th percentile. For simplicity of interpretation, both ν and *Y* are multiplied by 100, so a one unit change in each represents a 1% change in time.

The quasi-structural model yields very precise estimates that affirm the relevance (non-zero value of) all three structural parameters (standard errors in parentheses): ν_{EX} = -2.99 (0.001), *k1* = 1.05 (0.01) , *k2* = 4.67 (0.02). The estimate of ν_{EX} places the extensive margin at 3%, or 24 minutes, off-pace to break twenty-four hours; the other estimates have no natural interpretation. The incentives in the WS100 are sufficiently strong that the structural estimates are also precise: **P** = 0.17 (0.03); γ = 0.03 (0.002); σ_ϵ = 1.32 (0.05). All three are reasonable. The estimate of σ_ϵ suggests a mild degree of difficulty regulating one's pace, such that a finish time deviation of one percent from expected is not uncommon. The estimate of the fatigue parameter γ is positive but small: the 3% reduction in times observed by runners at the extensive margin increases the marginal cost of effort by about 10%. The **P** estimate indicates the value of passing the threshold is five times greater than the costs incurred to lower one's time by one percent. This is sufficient motivation to induce substantial effort. The ability-effort profile implied by these estimates, given by the dashed line in Figure 7a, is reasonably consonant with the semiparametric profile, suggesting that our simple model does an adequate job representing the major forces at work.

In the market-based theoretical model above, these structural estimates would allow the effect of the threshold on effort and efficiency to be determined. Given the very large deviation in performance between passers and nonpassers, the estimate of **p** in this market would be 0.0072, which is insufficient to motivate additional effort. This situation falls in the range of potency identified in the Appendix and in the lowest shaded band in Figure 3: actual effort under direct measurement, like the efficient level of effort, is zero, so the threshold inefficiently motivates greater

effort than direct measurement does. The accuracy of information clearly diminishes under the threshold as well.

For the WS100, instead, these structural estimates reveal the limits of human motivation. Like the semiparametric estimates, they imply that runners who are not motivated by the threshold finish well below their potential. But what governs the location of the line separating these runners from those who are so motivated–the extensive margin? In a five-kilometer race, the primary factor limiting threshold incentive effects would be the rapid buildup of lactic acid during the kick. There the fatigue parameter would be large, implying that it would be hard to kick much faster, and the location of the extensive margin would be determined mostly by physiological limits in the muscles and circulatory system, consistent with the "cardiovascular/anaerobic/catastrophe" model of fatigue.

Here, in contrast, the low value of the fatigue parameter suggests different mechanisms are at work. One is physiological: the "central governor" model of fatigue, which suits ultramarathoning, operating through the central nervous system, and supports our low fatigue estimate. (These two fatigue models are discussed and compared, in relation to distance running, by Weir et al., 2006; see also Millet, 2011. See Hamilton, 2013, on the significant role of "mental toughness" in ultramarathon performance.) But the other is preference-based, as belied by the high success rate of those runners who try to attain the twenty-four hour threshold. Figure 6b indicates that runners slightly above the extensive margin stand a more than even chance of passing the threshold. Runners slightly below this margin could expend the same amount of additional effort as those slightly above, and still stand a reasonable chance of succeeding. This is not a matter of limits, but a matter of choice.

The structural model indicates that even modest increases in runners' desire to achieve the threshold, **P**, would generate notable changes in behavior. When **P** increases from 0.17 to 0.20, for

example, the extensive margin moves back one percentage point and maximum effort, which occurs at that margin, increases by one percentage point. This is the ultimate irony of the Western States 100: in one of the toughest endurance races in the world, most finishers choose not to "use up all the gas in the tank."

IV. Conclusion.

Threshold-based incentives provide a rich arena for examining basic precepts of economic theory, yielding a multifaceted set of unusual implications that can be tested structurally or with non-restrictive semiparametric methods. These methods can uncover the positive and normative effects of these types of incentives in many areas of economic life. In our application to ultramarathoning, they show that, while it may be seemingly irrational to focus on an arbitrary time threshold of twenty-four hours, attempts to meet this threshold are anything but irrational, reflecting a clear, deliberate strategy that slowly unfolds throughout the race and closely conforms to the predictions of theory.

APPENDIX

Derivation of Effort under Direct Measurement and a Threshold. Given the cost function $C(f) = \exp(\gamma f) - 1$, the log of marginal costs is $\ln(\gamma) + \gamma f$. Similarly, the expected marginal benefits of effort, $\mathbf{P}\Phi'(\bullet)/\sigma_\epsilon$, logged, equal $\ln(.4\mathbf{P}/\sigma_\epsilon) - (f+\nu)^2/\sigma_\epsilon^2$. Given these, the program and solution for effort under three systems of measurement and reward is as follows.

- Direct, Perfect Measurement: max $\mathbf{p}f - C(f)$; $f_{\text{PERFECT}} = (1/\gamma)\ln(\mathbf{p}/\gamma)$.

- Direct, Imperfect Measurement: max $\mathbf{p}\sigma_\nu^2/(\sigma_\nu^2+\sigma_\epsilon^2)f - C(f)$;
  $f_{\text{IMPERFECT}} = (1/\gamma)[\ln(\mathbf{p}/\gamma)+\ln(\sigma_\nu^2/(\sigma_\nu^2+\sigma_\epsilon^2))] = f_{\text{PERFECT}} - (1/\gamma)\ln(1+\sigma_\epsilon^2/\sigma_\nu^2)$.

- Imperfect Measurement, with a Threshold Placed at Zero: max $\mathbf{P}\Phi(\nu+f) - C(f)$;
  $f_{\text{THRESHOLD}} = -(\gamma\sigma_\epsilon^2 + \nu) + (\gamma^2\sigma_\epsilon^4 + 2\gamma\sigma_\epsilon^2\nu + 2\sigma_\epsilon^2\ln(0.4\mathbf{P}/\gamma\sigma_\epsilon))^{\wedge}½$.

Proofs of Five Behavioral Properties. When positive, $f(\nu)$ in equation **(1)** satisfies the following conditions: C1) $f' > -1$; C2) $f'' < 0$; C3) $f''' > 0$; and C4) $f' = 0$ implies $f = -\nu$ and $y = 0$.

Let $\nu^* = \text{argmax}\, f(\nu)$. C4 (along with C1, when the maximum is at the extensive margin) ensure $\max(f) \geq -\nu^*$, so that $y(\nu^*) \geq 0$. These individuals' chances of passing the threshold are at least 50%, proving the Peak Proximity Property. And $\max(f) \geq -\nu^*$, along with C1, ensures $f(0) > 0$, the Precautionary Effort Property.

The Sawtooth Property is trivial if ν^* occurs at the extensive margin. For interior maxima, along with the presence of the extensive margin, $f'(\nu^*-d) > -f'(\nu^*+d)$ for any $d > 0$ by $f' = \int f''$ and C3. This Property, along with C1 and $y = f + \nu$, ensures the Stair Step Property.

Because f=0 below the extensive margin, while $f(0) > 0$, C2 and C4 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.

Motivational Effects. For all $\{\sigma_\epsilon, \sigma_\nu\}$, $f_{\text{IMPERFECT}} = [\ln(\mathbf{p}/\gamma) - \ln(1+\sigma_\epsilon^2/\sigma_\nu^2)] / \gamma$ is nil when $\mathbf{p}/\gamma \leq (1+\sigma_\epsilon^2/\sigma_\nu^2)$. The efficient level of effort, $f_{\text{PERFECT}} = \ln(\mathbf{p}/\gamma) / \gamma$, is nil when $\mathbf{p}/\gamma \leq 1$.

For a threshold, derivation of equation **(1)** shows that maximum effort never exceeds $\ln(.4\mathbf{P}/\gamma\sigma_\epsilon)/\gamma = \ln(.4\Delta\bar{y}\mathbf{p}/\gamma\sigma_\epsilon)/\gamma$. One can also show that $E(y|Y>0,f=0) = 0.8\sigma_\nu\bullet[\sigma_\nu/(\sigma_\nu^2+\sigma_\epsilon^2)^{½}]$. Thus, in the absence of effort, $\Delta\bar{y} = 1.6\sigma_\nu/(1+\sigma_\epsilon^2/\sigma_\nu^2)^{½}$, and threshold effort is nil when $\mathbf{p}/\gamma \leq 1.56(1+\sigma_\epsilon^2/\sigma_\nu^2)^{½}\sigma_\epsilon/\sigma_\nu$. Above this value, one can be sure that $f > 0$ for ν sufficiently close to 0.

Therefore, threshold effort will be positive, and effort under direct measurement nil, when $1.56(1+\sigma_\epsilon^2/\sigma_\nu^2)^{½}\sigma_\epsilon/\sigma_\nu < \mathbf{p}/\gamma \leq 1+\sigma_\epsilon^2/\sigma_\nu^2$. One can easily show the set of parameters satisfying these criteria is non-empty whenever $1.2\sigma_\epsilon < \sigma_\nu$. Mean threshold effort will continue to exceed effort under direct measurement for some (not necessarily small) range of $\mathbf{p}/\gamma$ exceeding $1+\sigma_\epsilon^2/\sigma_\nu^2$.

Using similar logic, one can show that threshold effort will be positive, and efficient effort zero, when $1.56(1+\sigma_\epsilon^2/\sigma_\nu^2)^{½}\sigma_\epsilon/\sigma_\nu < \mathbf{p}/\gamma \leq 1$. The set of parameters satisfying these criteria is non-empty whenever $1.8\sigma_\epsilon < \sigma_\nu$. When this condition holds, there is a set of parameter values for which threshold effort is inefficiently overprovided. When $1.2\sigma_\epsilon < \sigma_\nu < 1.8\sigma_\epsilon$, in contrast, there is a set of parameter values for which thresholds increase both effort and efficiency. When $1.2\sigma_\epsilon > \sigma_\nu$, then the effects of thresholds on average effort and efficiency cannot be characterized analytically, but simulations (as in Figure 4a) indicate that thresholds can still increase both effort and efficiency.

Signing Incentive Effects. Threshold effort $f(\nu)$ and outcomes $y(\nu)$ satisfy the following condition, which equates marginal costs and expected marginal benefits, in logarithms: $\gamma f = \ln(\mathbf{P}) - \ln(\gamma) + \ln(\Phi'(y)/\sigma_\epsilon)$. This, $\mathbf{P} = (\bar{y}_{PASSERS} - \bar{y}_{NONPASSERS})\mathbf{p} = \Delta\bar{y}\mathbf{p}$, and $\gamma f_{IMPERFECT} = \ln(\mathbf{p}/\gamma) + \ln(\sigma_\nu^2/\sigma_\nu^2+\sigma_\epsilon^2)$ yield: $\gamma(f(\nu) - f_{IMPERFECT}) = \ln(\Delta\bar{y}) + \ln(1+\sigma_\epsilon^2/\sigma_\nu^2) + \ln(\Phi'(y)) = \ln(\Delta\bar{y}) + \ln(1+\sigma_\epsilon^2/\sigma_\nu^2) - \ln(2.5\sigma_\epsilon) - y(\nu)^2/2\sigma_\epsilon^2$. Taking expectations across ν, and using $\text{var}(Y) = \text{var}(y) + \sigma_\epsilon^2$, yields:

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

Given σ_ϵ, all right-hand side terms can be inferred from the data, so the net incentive effect is signed.

Weak Identification in Structural Estimation. Equation **(1)** implies two relationships needed for identification of the structural parameters: $\gamma\sigma_\epsilon^2 = k1$, $\ln(.4\mathbf{P}/\gamma\sigma_\epsilon)/\gamma = k2$, where $k1$ and $k2$ are constants that are easily expressed, for example, as functions of the horizontal and vertical intercepts.

The condition satisfied by the extensive margin, $\{\nu_{EX}, f_{EX}\}$, is $\mathbf{P}[\Phi((\nu_{EX}+f_{EX})/\sigma_\epsilon) - \Phi(\nu_{EX}/\sigma_\epsilon)] = C(f_{EX}) - C(0)$, where $C(f) = \exp(\gamma f) - 1$ is the cost function. Define $\phi(\bullet;\sigma_\epsilon)$ as the normal density function with mean zero and standard deviation σ_ϵ. Then, replace $\Phi(\nu_{EX}/\sigma_\epsilon)$ with its second order Taylor series approximation around $\nu_{EX}+f_{EX}$, and replace $C(0)$ with its second order Taylor series approximation around f_{EX}. Then, to the second order, $\{\nu_{EX}, f_{EX}\}$ satisfies:

$$f_{EX}\mathbf{P}\phi(\nu_{EX}+f_{EX};\sigma_\epsilon) - f_{EX}^2\mathbf{P}\phi'(\nu_{EX}+f_{EX};\sigma_\epsilon)/2 \approx f_{EX}C'(f_{EX}) - f_{EX}^2C''(f_{EX})/2$$

The first term on the left-hand side equals the first term on the right-hand side; this is simply the first order condition for optimality, $\mathbf{P}\phi = C'$, multiplied by f_{EX}. Furthermore, it is straightforward to show that $d\phi/dx = -x\phi/\sigma_\epsilon^2$ and $C'' = \gamma C'$. Eliminating the first term on each side and making these substitutions in the second term on each side yields:

$$f_{EX}^2\mathbf{P}y_{EX}\phi(\nu_{EX}+f_{EX};\sigma_\epsilon)/2\sigma_\epsilon^2 \approx -f_{EX}^2\gamma C'(f_{EX})/2$$

Again using $\mathbf{P}\phi = C'$ and rearranging yields:

$$-(\nu_{EX}+f_{EX}) = -y_{EX} \approx \gamma\sigma_\epsilon^2 \equiv k1$$

Thus, when higher order effects are small, $k1 \approx -y_{EX}$. Substituting this relation into equation **(6)** solves for ν_{EX} in terms of $\{k1, k2\}$: the three structural parameters are weakly identified. As $\gamma\sigma_\epsilon^2 > 0$, this is most likely to happen when there is an interior maximum, so that y_{EX} is negative, or when $0 \approx y_{EX} << f_{EX}$, as in Figure 4. By the geometry of Figure 1, this will occur when incentives are not too potent–that is, when the slope of the marginal costs line, γ, is sufficiently large relative to the height of the parabola of marginal benefits, $\mathbf{P}/\sigma_\epsilon$.

REFERENCES

Allen, Eric, Patricia Dechow, Devin Pope, and George Wu. "Reference-Dependent Preferences: Evidence from Marathon Runners," NBER Working Paper 20343 (2014).

Baghestanian, Sascha, and Sergey Popov. "On Publication, Refereeing, and Working Hard." Manuscript, Queen's University, Belfast (2014).

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

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

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

Chakrabarti, Rajashri. "Vouchers, Public School Response, and the Role of Incentives: Evidence from Florida," *Economic Inquiry* 51:500-526 (2013).

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).

Dubey, Pradeep, and John Geanakoplos. "Grading Exams: 100, 99, 98... or A, B, C?" *Games and Economic Behavior*, 68:72-94 (2010).

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," *Empirical Economics*, 44:1563-1592 (2013).

Grundfest, Joseph A., and Nadya Malenko. "Quadrophobia: Strategic Rounding of EPS Data," Rock Center for Corporate Governance at Stanford University Working Paper No. 65 (2011).

Hayfield, Tristen, and Jeffrey Racine. "The np Package." Manuscript (2011).

Hamilton, Michelle. "How Much Does Mental Toughness Affect Race Times?" www.runnersworld.com/sports-psychology/how-much-does-mental-toughness-affect-race-times (2013).

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

Henderson, Daniel, and Christopher Parmeter. *Applied Nonparametric Econometrics*. Cambridge: Cambridge University Press (2014).

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

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).

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

Millet, Guillaume. "Can Neuromuscular Fatigue Explain Running Strategies and Performance in Ultra-Marathons?" *Sports Medicine* 41:489-506 (2011).

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).

Pagan, Adrian, and Aman Ullah. *Nonparametric Econometrics*. Cambridge University Press (1999).

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).

Stanley, Ted, and Hristos Doucouliagos. *Meta-Regression Analysis in Economics and Business*. Oxford: Routledge (2012).

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

Wang, Li, and Lijian Yang. "Spline Estimation of Single-Index Models," *Statistica Sinica* 19:765-783 (2009).

Weir, J., T. Beck, J. Cramer, and T. Housh. "Is Fatigue All in Your Head? A Critical Review of the Central Governor Model," *British Journal of Sports Medicine* 40:573-588 (2006).

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

Yim, Andrew. "Mixture and Continuous 'Discontinuity' Hypotheses: An Earnings Management Model with Auditor-Required Adjustment," SSRN Working Paper (2013).

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

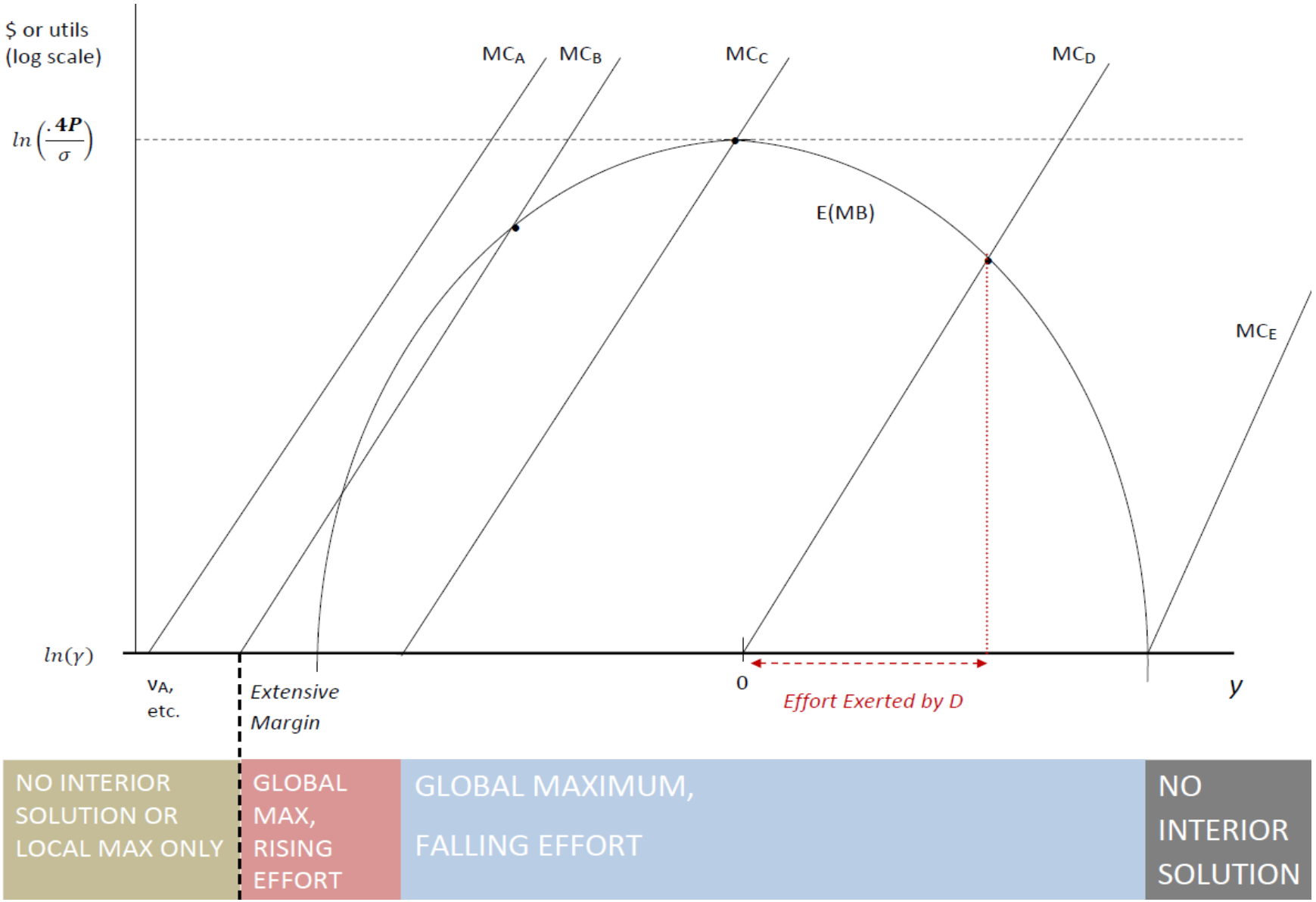


Figure 2. (a) Ability-Effort Locus, (b) Ability-Performance Locus.

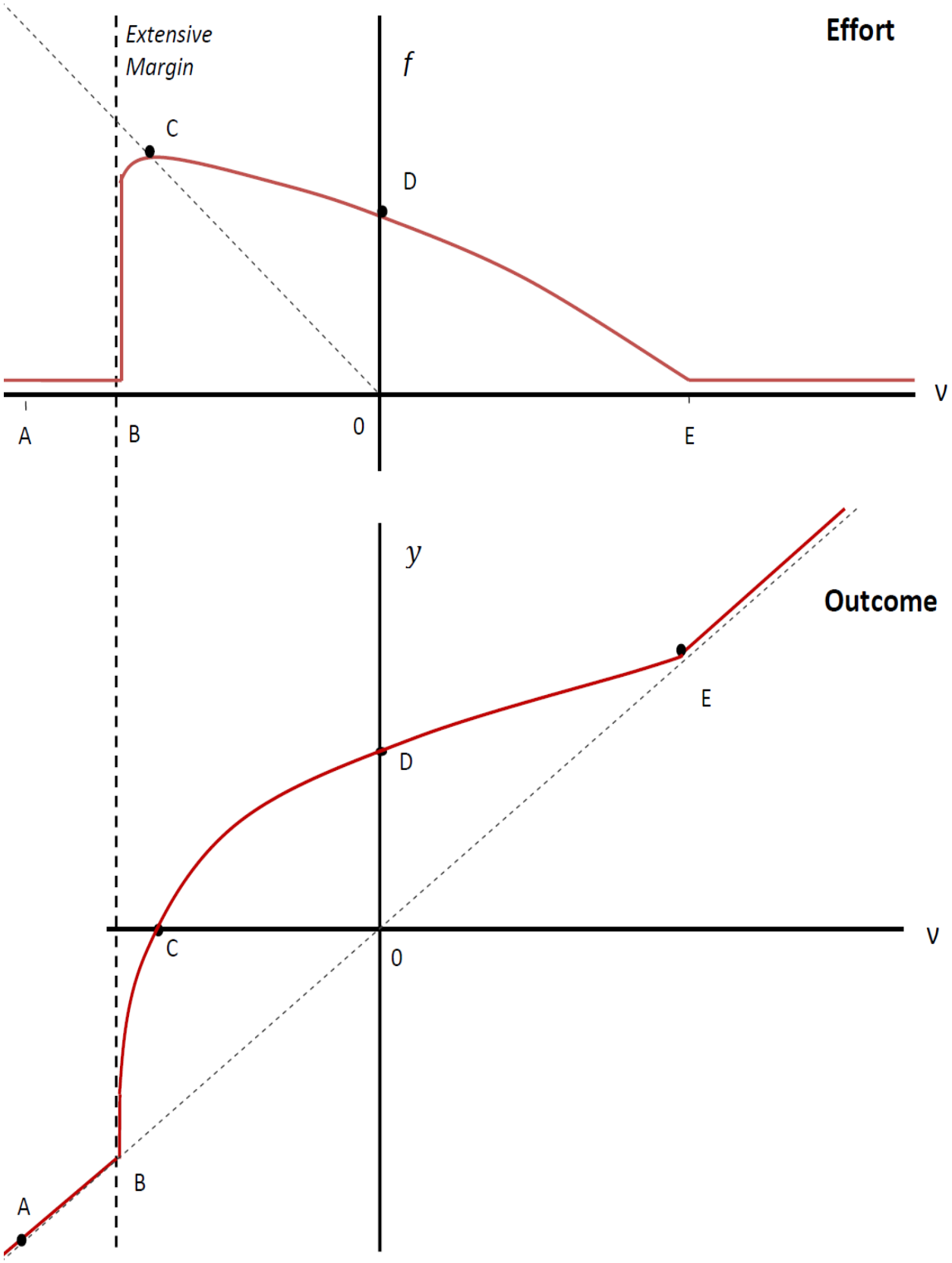
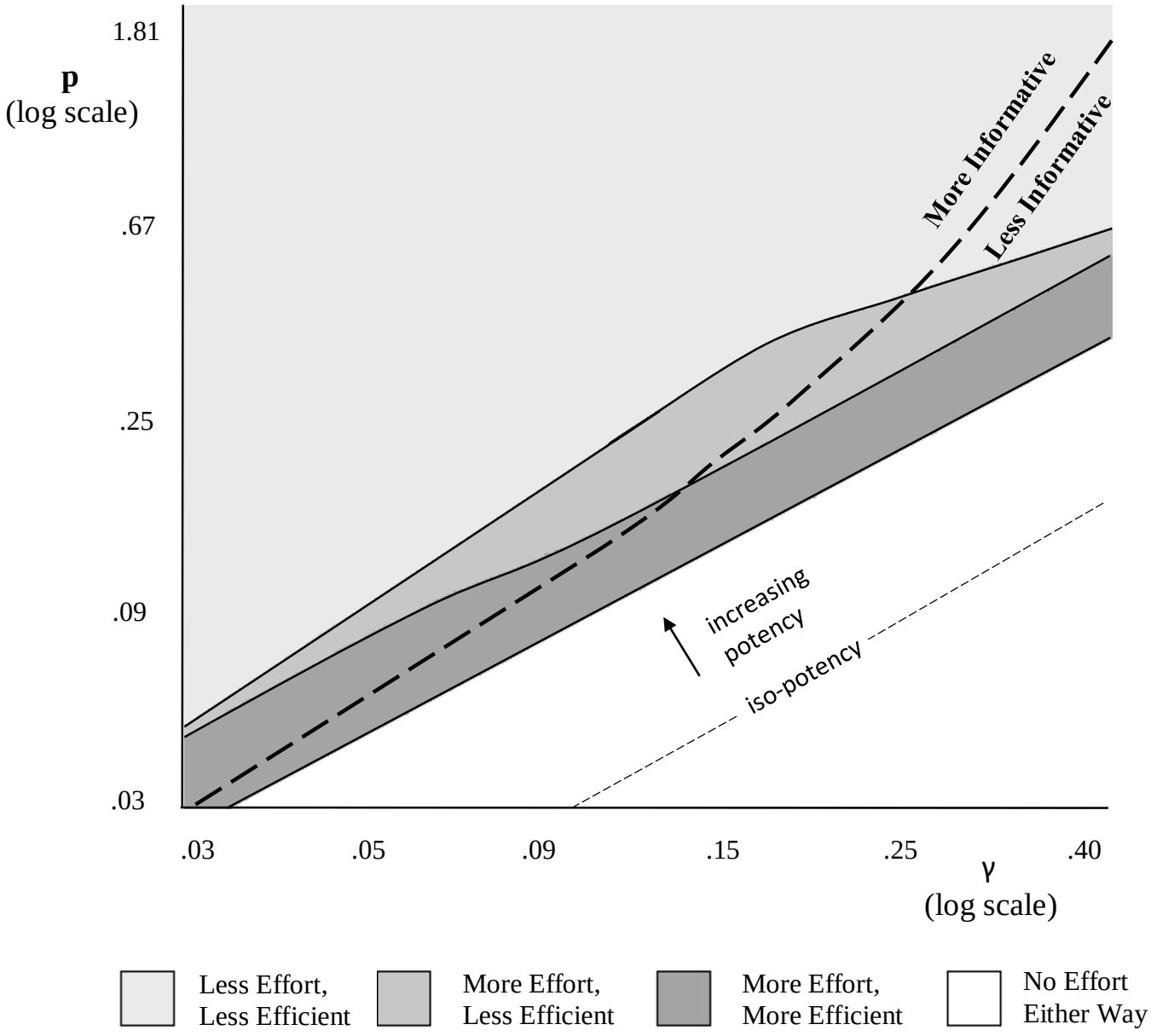
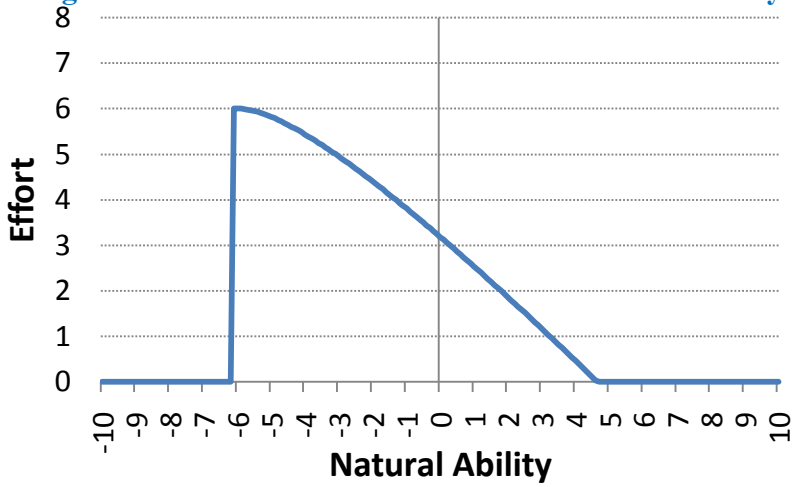


Figure 3. Mean Effort, Efficiency, and Information Accuracy under a Threshold, Compared to Direct Measurement. (Drawing is to scale. The term σ_ε is normalized to one, and $\sigma_\varepsilon/\sigma_v = ½$.)

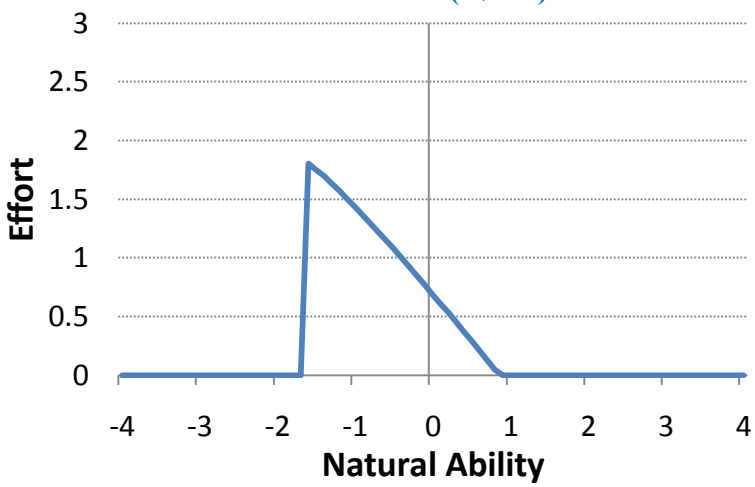


Note: Information accuracy is determined by the standard deviation of true performance, *y*. The standard deviation of *y*, conditional on passing, under the threshold is compared to the standard deviation of *y*, conditional on observed *Y*, under direct measurement. If the former is smaller (larger), the threshold is more (less) informative than direct measurement. The threshold is located one standard deviation of the error (σ_ε) above mean ability. To a very close approximation, this is where thresholds' motivational properties are strongest. The figure is invariant to proportional changes in all parameters (p, γ, σ_ε, and σ_v).

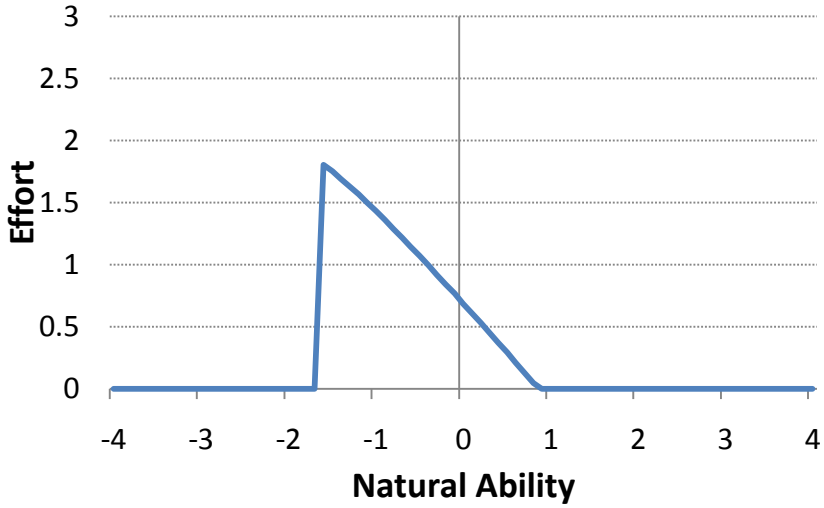
Figure 4: Theoretical Relation between Natural Ability and Effort under a Threshold Placed at Zero ($\sigma_\nu = 3$)



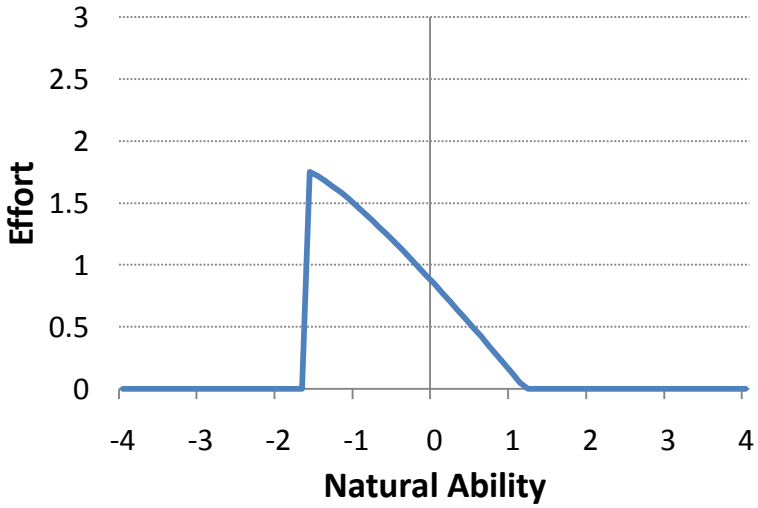
(a) Parameter Values: $\gamma = 0.2$, $\sigma_\varepsilon = 3$, **P** = 5, mean $\nu = -3$, **p**=0.71.
$f_{\text{PERFECT}} = 6.34$, $f_{\text{IMPERFECT}} = 2.70$, $f_{\text{THRESHOLD}} = 3.73$.



(b) Parameter Values: $\gamma = 1.2$, $\sigma_\varepsilon = 0.4$, **P** = 12, mean $\nu = -1$, **p**=2.25.
$f_{\text{PERFECT}} = 0.53$, $f_{\text{IMPERFECT}} = 0.51$, $f_{\text{THRESHOLD}} = 0.32$.



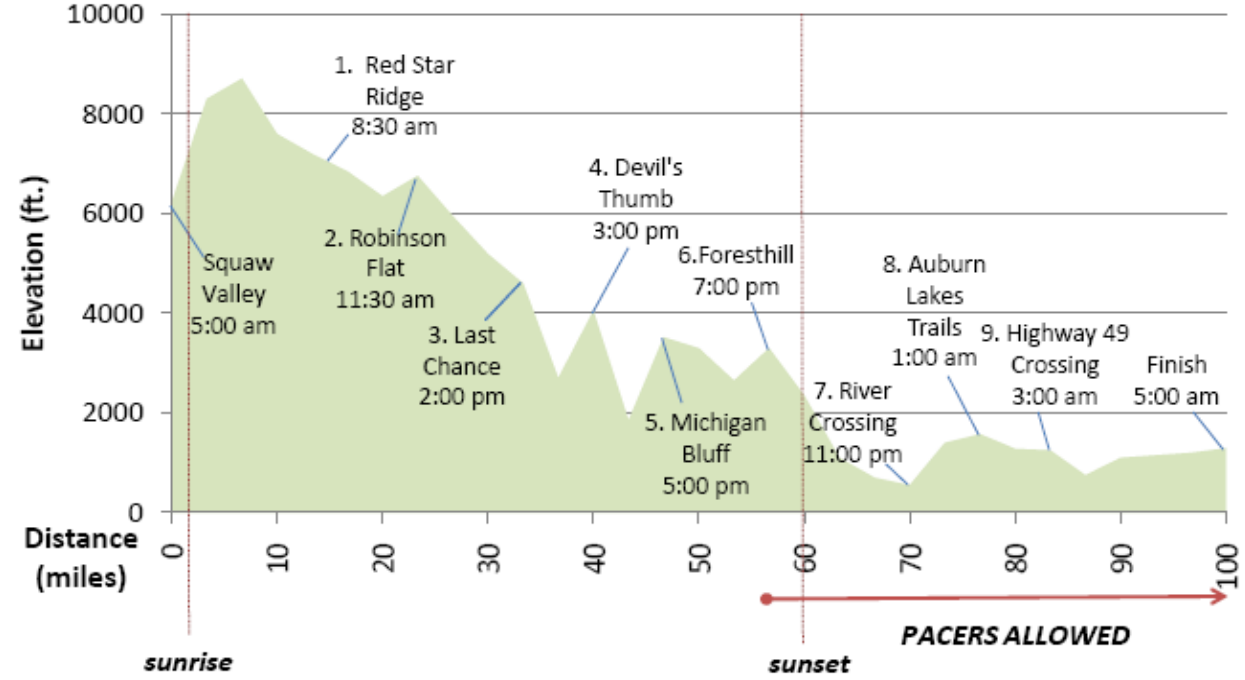
(c) Parameter Values: $\gamma = 0.1$, $\sigma_\varepsilon = 2$, **P** = .6, mean $\nu = -2$, **p**=0.13.
$f_{\text{PERFECT}} = 2.75$, $f_{\text{IMPERFECT}} = 0$, $f_{\text{THRESHOLD}} = 0.31$.



(d) Parameter Values: $\gamma = 0.6$, $\sigma_\varepsilon = 0.8$, **P** = 3½, mean $\nu = -1$, **p**=0.66.
$f_{\text{PERFECT}} = 0.16$, $f_{\text{IMPERFECT}} = 0.04$, $f_{\text{THRESHOLD}} = 0.35$.

Note: Given **P**, mean *y* is calculated for passers and nonpassers across the full domain of ν. From this **p** is backed out and used to calculate f_{PERFECT} and $f_{\text{IMPERFECT}}$, which are independent of ν; $f_{\text{THRESHOLD}}$ refers to mean effort across all agents in the presence of the threshold.

Figure 5. Course Layout and Time Distributions at Splits 2, 6, and 7 and the Finish, Western States 100. (Times in the Course Layout are for a runner who finishes in twenty-four hours.)



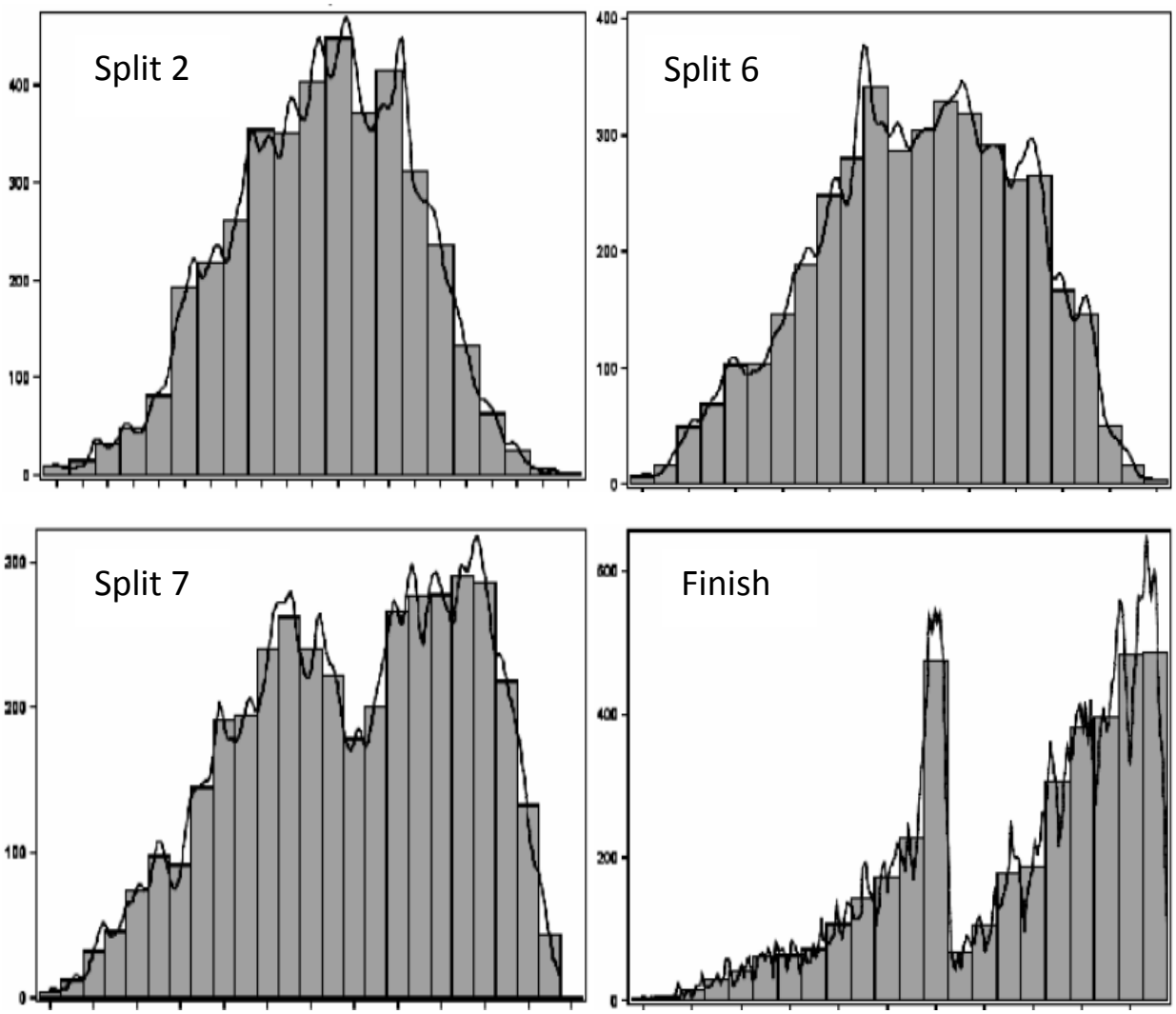
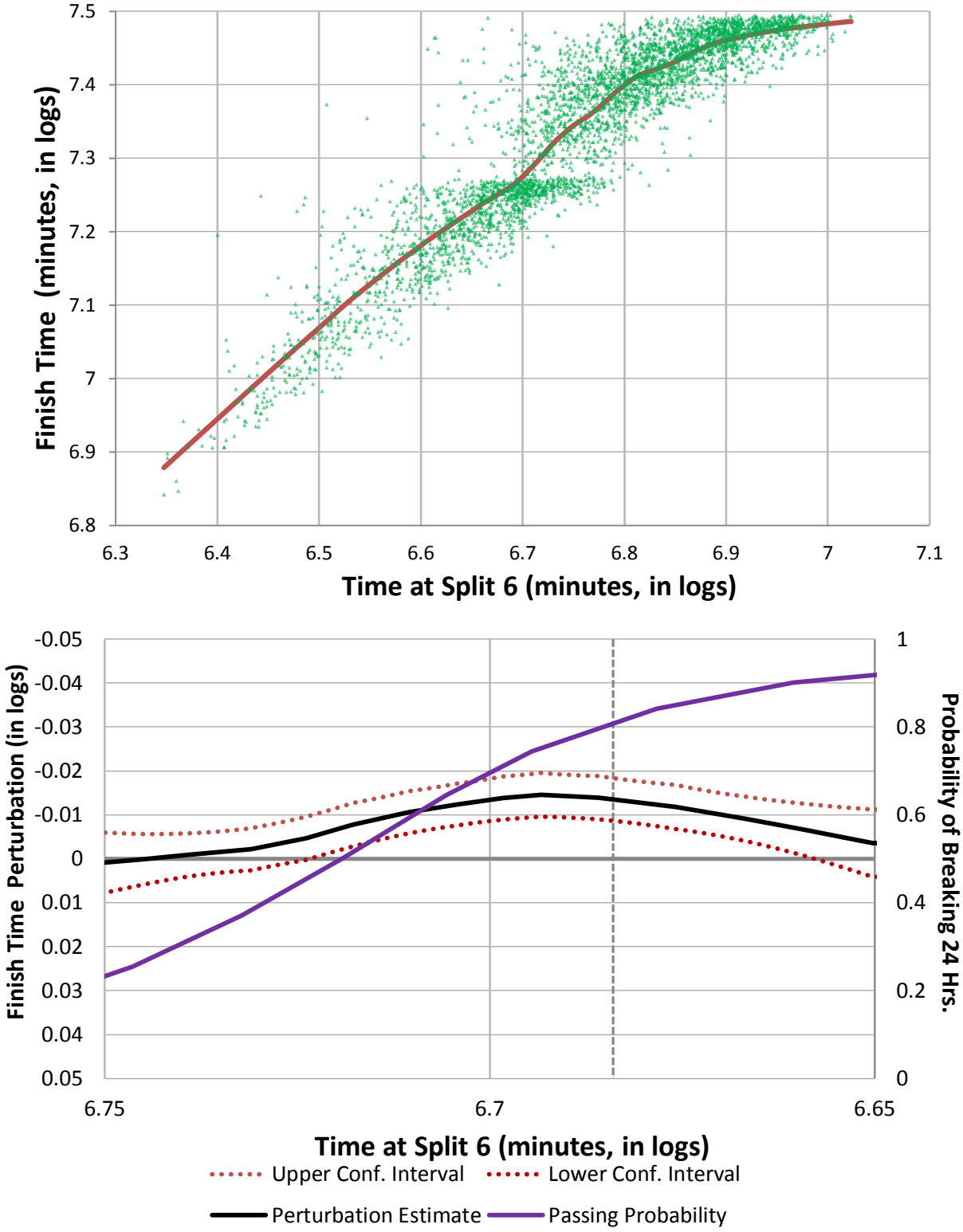
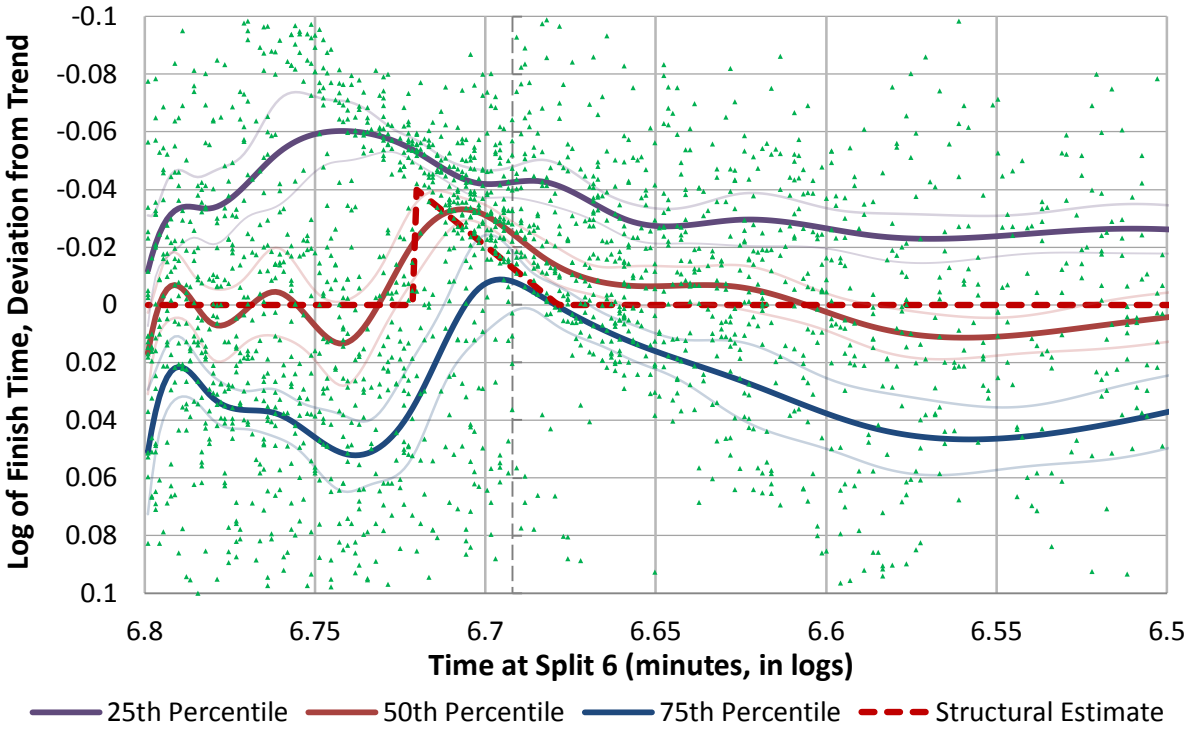


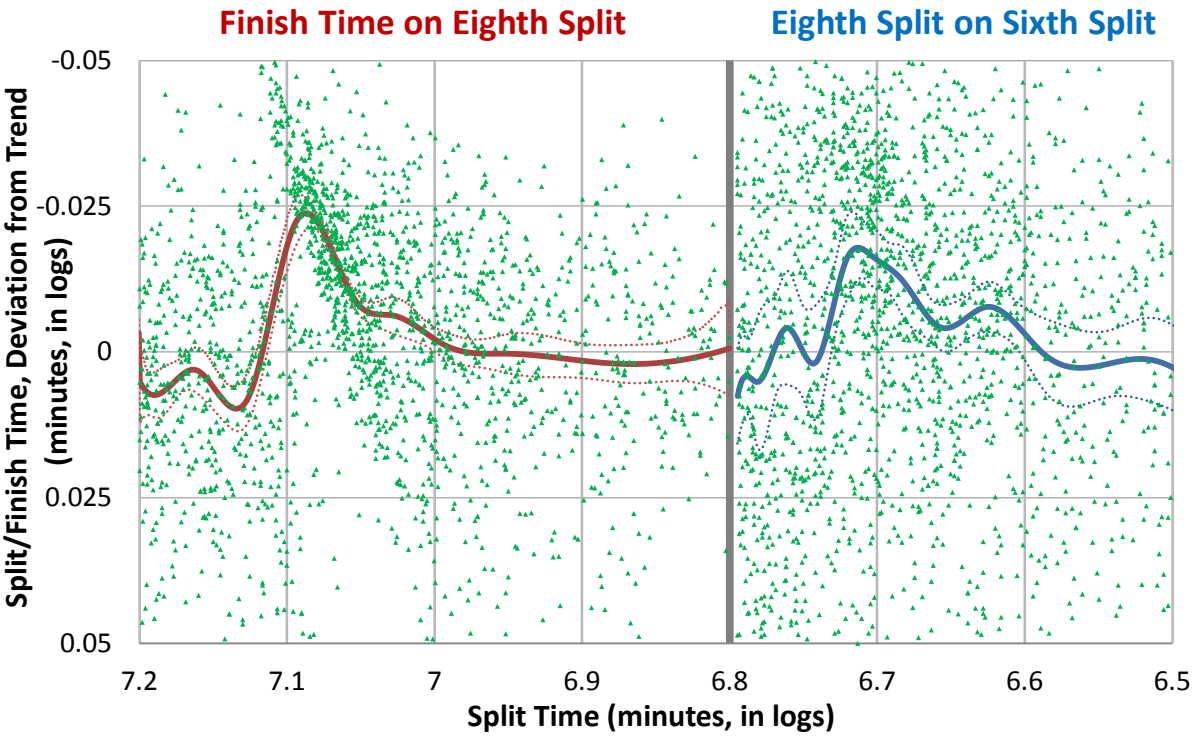
Figure 6. (a) Split 6 and Finish Time Scatterplot, with Smoothed Mean, (b) Effort Perturbation.



Note: The dashed vertical line corresponds to a predicted finish time, *absent the effort perturbation*, of 24 hours.

Figure 7. Detrended Smoothed Quantiles: (a) Finish Time on Sixth Split, (b) Inter-Split Breakdown.





Note: In both graphs, dots represent individual time deviations from the parametric trend. In Figure b, quantiles are estimated at the 50th percentile. 95% confidence intervals included.

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), Yim (2013), and more	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	McEwan and Salti-banez (2005), Reback (2008), Chakrabarti (2013), Grant and Green (2013), 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	school districts focus their efforts on those students who are near the border between passing and failing standardized tests, improving the rate at which those students pass the tests
analyst / publication bias in several fields of social science	Card and Krueger (1998), Tufte (2006), Gerber and Malhotra (2008), Stanley and Doucouliagos (2012)	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