On the Theory of Product Market Characteristics and the Industry Life Cycle

Kenneth L. Simons

Department of Economics
Rensselaer Polytechnic Institute
110 8th Street
Troy, NY 12180-3590
United States
Tel.: 1 518 276 3296
Email: simonk@rpi.edu
Web: www.rpi.edu/~simonk

April 3, 2006

On the Theory of Product Market Characteristics and the Industry Life Cycle

Abstract:

Does industry competition depend largely on product-specific traits? If so, what traits matter how? A model is developed in which the nature of technological opportunities in an industry affects industry evolution over long periods following the inception of a product. Continuous-time firm decisions dictated by optimal control theory provide a mathematical basis for proofs of inter-firm differences and industry outcomes. Analysis of the model shows that, with firms optimizing discounted profit streams by choosing entry, exit, growth, and research spending, alternative industry dynamics arise depending on the degree of relevant technological opportunity. The theory provides an explanation for why some industries experience shakeouts and concentration while others do not, and matches with rich empirical findings reported in a companion paper.

Keywords: industry dynamics, product life cycles, technology, optimal control, theory of industrial organization

JEL codes: L11, O33

On the Theory of Product Market Characteristics and the Industry Life Cycle

I. Introduction

In a new generation of research, economists have developed dynamic models of industry structure that explain features of real industries heretofore unexplained by static models of competition. Synchronous with the theoretical work, empirical research is documenting important patterns in industry dynamics. One of the central patterns studied has been industry shakeouts, in which a dropoff in the number of firms in an industry follows an initial rise in firm numbers. This paper extends previous theories of industry dynamics and shakeouts to explain important, distinctive patterns in industry outcomes as a result of rational firm behaviors.

The goal is to develop a theory that explains the broad long-term patterns in evolution of industry structure, and importantly, inter-industry differences in this evolution, for the majority of (non-declining) industries. The idea, an old idea in industrial organization economics, is that underlying industry traits drive the outcomes of different industries in predictable ways.[1] While Klepper (1996) characterized shakeouts as a stylized fact about industries and sought to explain this fact common across industries, the present paper emphasizes differences across industries. Inter-industry differences in competitive dynamics are explained as outcomes of "product market characteristics," underlying parameters specific to particular industries. The parameter considered here, based on past findings suggesting that technology is extremely important in industry shakeouts, is a particular type of technological opportunity. Relevant technological opportunity allows research and engineering (R&E) specific to a product or its production process, excluding new-product development, and yields findings that cannot profitably be licensed. The degree of such technological opportunity will be referred to by a parameter alpha, similar in spirit to the alpha use by John Sutton (1998) to explain technology-based differences

[1] Schmalensee (1989) reviews a large empirical literatature on cross-sectional differences in industry structure. The review reflects the large base of relevant work, but ultimately also reflects the limits of the cross-sectional empirical approach. Schmalensee points to the need for theory to understand such cross-sectional differences. A further limitation of previous cross-industry comparisons is that, given available data, little evidence has available on cross-industry dynamics rather than comparisons at a single point in time.

in the lower bound to static industry concentration. Such a characterization, if correct, promises the potential to tailor industrial and competition policy, and perhaps even to make a priori forecasts, in ways that account for key differences in dynamic industry outcomes.

The underlying engine behind the dynamics developed here is only one of various possible choices. Why does the paper not focus for example on demand-side externalities, a topic of much interest amplified by recent investigations about the market dominance of Microsoft in operating systems, or buyer switching costs (Klemperer 1995), or production scale economies, or increasing returns in advertising (Sutton 1991)? The answer is that although these issues are important, explain important patterns in industries, and have been the subject of excellent theoretical and empirical research, they have proved less suitable to explain the main long-run dynamics of industry structure in the majority of industries, judging from previous empirical work.[2,3] That previous work suggests that differences in technological innovation have been central to fueling the market dominance that emerged in industries with shakeouts.

---

[2] Consider for example network externalities, unquestionably an important topic and the subject of much excellent work. Network externalities yield a process of concentration and market dominance in some industries, but the set of such industries is tightly circumscribed; consider the most cited examples. Microsoft became the dominant producer of operating systems and of some other types of software, but this dominance was constrained to specific industries and did not extend to the manufacture of computers. The VHS format with its superior recording time beat out Sony's Betamax, and the industries manufacturing videocassette recorders and videocassettes remained the province of many manufacturers (Liebowitz and Margolis, 1995). The QWERTY typewriter keyboard layout became the dominant layout, but this did not restrict the typewriter industry's manufacturers, virtually all of which freely and quickly adopted the design. While there are many cases in which network externalities matter, in few industries did they impact the success and failure of companies so strongly as to have yielded concentration.

[3] Production scale economies appear typically to play a limited role in yielding concentration, judging from previous studies that examine the engineering issues and equipment involved in specific industries. This assessment may be consistent with Alfred D. Chandler's (1990) landmark history of twentieth century industrial enterprise, in which Chandler identifies the contribution of company size to commercial success as stemming not primarily from its

The theory developed here yields implications that are empirically testable and *distinctive*. Various alternative theories suggest contrasting patterns in early mover advantage, exit, and the role of technological innovation, so that empirical evidence can be used to probe which mechanism(s) appear to be at work in industry dynamics. Appropriate cross-national comparisons provide a means to gauge whether underlying product market characteristics, rather than nationwide institutions or random outcomes, are the leading systematic explanators of industry outcomes. Variations across industries in early entrant advantages, considered both by the literature on entry timing in economics and by a vast empirical literature in management, are argued to stem from a technology-driven process of competition, tied up with the degree of shakeout, again in an empirically testable manner.

The model characterizes firm behavior and competition in a new product market in a very general manner. Differences between firms at any point in time are completely characterized by two traits: the skill of employees at research and engineering, and the quantity of output. Skill is assumed to be a fixed trait, because it is difficult for a firm to attract more skilled personnel if the firm does not have the advantages that skill implies. In contrast, quantity of output can change over time, but as Penrose (1959) characterizes, rapid growth is especially expensive because it requires a large fraction of managerial time to train new employees and acquire and install plant and equipment.

To a varying extent depending on product- and market-specific traits, larger firms may have relatively high profit margins, and the model represents their profit advantage as stemming from research and engineering. Within a product market, profit advantage does not stem from development of new products outside the market, and hence the term R&E rather than R&D is used to emphasize this difference compared to the firm's overall research and development. Firms may create new variants of the existing product, so long as they satisfy the same consumer need as the original product, and such variants are characterized here as quality changes. Firms may also improve efficiency of the manufacturing process. The two types of innovation are

---

production equipment but from the scale and scope of the organization and its managerial activity. The research and engineering related economies considered here are a type of scale economy and can be analyzed using traditional static models of scale economy, but this would shed little light on the dynamics of industries.

treated together as affecting cost per unit of quality. The lower is a firm's cost per unit of quality, the greater is its price-cost margin and hence its profitability. Mathematically this effect is embodied through cost rather than price, but the idea is that quality improvements enhance the price a firm can charge and hence enhance profit margins in the same way that cost improvements enhance profit margins. The degree of improvement in unit cost per unit of quality depends on firms' R&E budgets, so that larger firms reap greater gains from R&E because their total cost is reduced for every unit they produce; this is similar to earlier treatments including Flaherty (1980) and Shaked and Sutton (1987).

The model builds on earlier models of technology and market structure such as Nelson and Winter (1978), Dasgupta and Stiglitz (1980), Shaked and Sutton (1987), and Sutton (1998). Most immediately the model builds on Klepper's (1996) model of industry dynamics, with three key differences. First, the present model explicitly characterizes inter-industry differences in the potential for within-firm R&E. The R&E relevant for inter-industry differences is difficult to embody in saleable goods or to license effectively, and hence is not marketed between firms. (Saleable R&E outputs have a different character in that the seller may price the good proportionally to the user's financial returns and thus eliminate any advantage to particular using firms.)

Second, this model excludes development R&D and deliberately treats process and product R&E as comparable. In contrast, Klepper (1996) characterizes product R&D separately in order to generate a shift in importance from product to process R&D over the industry life cycle, which Abernathy and Utterback (1978; Utterback and Abernathy, 1975) depicted as common. Such a product-process shift has not generally been borne out in empirical studies, which might better be characterized as showing the continuing importance of both types of innovation (McGahan and Silverman, 2001).

Third, firms are assumed to maximize not current-period profit, but long-run discounted profit streams. Klepper's (1996) model assumes only current-period profit maximization, and it has been unclear whether the outcomes of that model would continue to hold with long-run optimization by firms. As is shown here, in fact, key outcomes are robust to profit maximization with any discount rate. Actual firms may not perfectly optimize, but even if they stray far from optimality, the forces at work in the model are powerful and consistent enough to drive a rich-

get-richer process, in which smaller and less-skilled firms are steadily driven out of business in high-R&E industries while firms remain relatively equal in low-R&E industries.

The model deliberately abstracts away from competitive oligopoly games and focuses on competition among many firms, each unable to affect market prices or the behavior of other firms. There are three justifications for this focus, beyond simple tractability of the analytic model. The intent is to analyze competition in the formative years of an industry, during which individual firms are initially fairly small relative to the total market, in comparison to possible long-run levels of concentration. Moreover, even in later years the effects of firms' oligopoly games may be unstable and in practice rather limited as long as more than a handful of substantial-sized producers remain, as is the case in most of the empirical studies discussed later in this paper.[4] And perhaps most importantly, it is difficult to know how to analyze or interpret game theoretic competitive behavior unless the baseline market behavior under perfect competition is understood. In the latter regard, the present model serves as a building block for possible future analyses involving competitive behavior.

The ultimate test of the theory is its fit with empirical reality. If the processes characterized here are leading drivers of industry outcomes, the implications of the theory should hold across a majority of industries despite the diverse forces at work in these industries. In a bi-national study of 18 industries over various eras of the twentieth century, Simons (2005) examines each of the predictions made here. The study is unique in that it relies on newly-collected evidence for many competitive-level industries (i.e. making products that are substantially substitutable for consumers) each with data spanning multiple countries and many decades from near the inception of a product, and each matched to firm-specific measures of technological change.

[4] One exception is UK tire manufacture, but even in this case in which a dominant UK producer held the majority of the market for an extended period, the attempts of manufacturers to collude appear to have had limited effect. Price wars in UK tire manufacture were common until at least the late 1930s, despite explicit attempts by manufacturers to set prices, and when these attempts succeeded they were soon broken up following regulatory proceedings (Monopolies and Restrictive Practices Commission, 1955).

The companion study's findings confirm each of the distinctive predictions derived herein. First, shakeouts occur to similar degrees in matched industries across countries, and even the timing of the shakeouts is highly correlated for matched pairs of industries across countries in which the industries developed. This finding suggests that underlying product market characteristics, not random forces or national idiosyncracies, drive processes of industry shakeout and concentration. Second, there is no consistent rise in the rate of exit associated with industry shakeouts, while a substantial fall in entry and consistently higher survival in the market of early than late entrants are associated only with industries that experience shakeouts. It appears therefore that shakeouts might typically be driven by a competitive process tied up with some advantage enjoyed by earlier entrants, such as the process described in the present theory, and in contrast that external causes such as technological changes do not systematically trigger shakeouts.[5] Third, in industries with strong shakeouts, early entrants have the greatest R&E output, and R&E output is most positively correlated with survival in the market for industries with strong shakeouts. This evidence is thus consistent with the idea that technological R&E opportunities underly the majority of cross-industry differences in changing firm numbers and (by implication) concentration. No doubt there are exceptions to the stylized theory presented here, but initial indications suggest the theory may aid understanding of key differences in industry outcomes.

[5] Klepper and Simons (2005) further assess the role of exogenous events versus continuous and presumably endogenous processes in industry shakeouts, and review related theories and empirical evidence.

II. The Model[6]

The model characterizes a product industry's competitive dynamic between firms. Firms are defined to be in the industry if they produce a particular product (which could be a service), which is first sold at time 0. Firms obtain positive expected returns by entering production of the product, and each continues to produce as long as production remains profitable. Dynamics arise among firms since two factors prevent an immediate jump to a static solution: knowledge diffusion constrains the number of potential entrants at each time, and growth costs constrain the pace of firms' expansion. Firms or potential firms with requisite skills and cap $P(t)$ ability to produce the product are limited in number because the information and resources necessary do not spread freely, and at any time those that could choose to enter production are termed potential entrants. New firms upon entry are producing 0 of the product *per diem*, and they pay to expand production over time by hiring and training employees and by buying and putting to use equipment, all costly processes that can take up an excessive share of a manager's time if expansion is rapid. Inter-firm differences arise from varying levels of two firm competencies: the skill of key personnel at management, research, and engineering (characterized as a fixed trait), and the size of the firm's production (a state variable). Firms' skill and size affect their incentives for and benefits from research and engineering (R&E) spending, and the resulting differences in R&E spending propel the most-skilled early entrants to outgrow later entrants and give them cost advantages that keep them profitable while higher-cost firms exit.

A. Potential Entrants and Skill

Potential entrants with the requisite knowledge and resources to enter the product market arise starting at time 0 at a (finite) density firms per year.[7] For potential entrant or firm i, its

[6] The following notations are used. A dot above the name of a variable indicates the total derivative of the variable with respect to time, for example $\dot{q}_i(t) = \frac{dq_i(t)}{dt}$. Two dots above the name of a variable indicate the second total derivative with respect to time. A prime indicates the derivative of a univariate function; for example $v'(u) = \frac{dv(u)}{du}$. The superscript –1 denotes the inverse of a function; for example $D^{-1}(Q;t)$ is the inverse demand function and gives the price at which the demand is Q at time t. An asterisk indicates an optimal value.

engineering and management personnel's skill is $s_i > 0$. Skill is fixed over time, because poorly skilled firms are unable to attract highly skilled employees.[8] Skill is distributed among new potential entrants, at each time t, with cumulative distribution function $H(s_i)$, which is a C^1 function strictly increasing from 0 at $\underline{s}$ to 1 at $\overline{s}$ for some finite lowest and highest possible skill levels $\underline{s}$ and $\overline{s}$. New potential entrants acquire the requisite knowledge and resources to enter at time E_i. They then choose whether to enter. They can enter any time from E_i onward, but it will turn out that they enter immediately if at all since the greatest benefit will stem from the earliest entry. Entrants initially have 0 output.

B. Expansion

After entry, firms determine their output $q_i(t) \geq 0$ at each t by choosing a time path of expansion. To expand or contract, the firm trains staff, reorganizes production, and purchases or disposes of equipment, yielding growth cost $g\left(\dot{q}_i(t)\right)$.[9] The cost $g(\cdot)$ is continuously

---

[7] The determinants of $P(t)$, such as knowledge diffusion processes, are fascinating to investigate in their own right, here the possible processes are kept completely general in order to maintain generalizable conclusions.

[8] A more precise model of the decisions of skilled employees to take jobs, and of the job offers of employers, would suggest some modification to numerical values but not to general trends in the conclusions. Larger less-skilled employers would have comparable incentive to hire skilled employees as smaller more-skilled employers, and under appropriate circumstances skilled personnel could have sufficient incentive to take on these new jobs. Walter Percy Chrysler's takeover of the ailing Maxwell Motor Corporation (recently merged with Chalmers) is an apparent example. This suggests that early entrants with moderately high skill may fare slightly better than predicted by the model through acquisition of more skilled employees.

[9] The identification of firm sizes through convex costs of growth dates back mathematically at least to M. Therese Flaherty (1980), and is apparent in Edith Penrose's (1959) discussion of the processes of firm growth. The cost of growth could more realistically be made a function $g\left(\dot{q}_i(t), q(t)\right)$ with $\frac{\partial g}{\partial q_i} < 0$ to allow the cost of growth to decline with firm size, with similar outcomes. This has not been shown here to avoid substantially more complicated mathematics.

differentiable and finite with a minimum at zero, such that $\frac{\partial g}{\partial \dot{q}_i}(0) = 0$ and $\frac{\partial^2 g}{\partial \dot{q}_i^2} > 0$; i.e., it is an increasing convex function of the absolute rate of expansion.

C. Research and Engineering

Firms lower production cost, and improve product quality thus raising consumer's willingness to pay for the good, through research and engineering (R&E). In the model the R&E benefits of higher price and lower cost are embodied for simplicity solely as cost reduction (price improvements under certain conditions have exactly analgous effects).[10] R&E spending is chosen at each point in time, yielding a time path of spending $r_i(t) \geq 0$. The R&E spending has an impact after a delay $\delta \geq 0$, when it lowers the average production cost of the good. Potential for R&E to reduce cost and improve quality is measured by the parameter $\alpha > 0$. Even with zero R&E, however, the average production cost a firm could have at time t is bounded by a maximum possible cost $\bar{c}(t)$. Firm R&E is often modeled using a quality ladder or learning curve, and the conceptual view is the same here, but the implementation is different in order to reduce dramatically the complexity of the resulting mathematics. Instead, R&E is portrayed as an independent decision and (eventual) outcome at each time t, and this portrayal will turn out to yield the same monotonic increase in knowledge stocks apparent in quality ladder and learning models, without affecting the implied nature of competition between firms.

Cost reduction due to R&E spending is a fraction of $\bar{c}(t)$. This fractional cost reduction is $f\left(\alpha s_i r_i(t-\delta)\right)$, with diminishing marginal returns such that $f(0) = 0$, $f' > 0$, and $f'' < 0$.[11,12] Cost reduction is greater for higher-skilled firms, since they are better at the R&E

[10] Price enhancements associated with greater willingness to pay for a firm's goods yield exactly analogous effects to cost reduction if they can be described by the same functional forms, which will involve decreasing marginal benefits. An extension to explicitly include product R&E is addressed later in the paper.

[11] In the model, $f(\cdot)$ declines solely with R&E spending at a time δ in the past, but in reality it would decline with R&E spending at a range of times. This simplification is a consequence of the approximation to quality ladder and learning curve models described above.

[12] Technically the domain of $f(\cdot)$ is defined to include positive infinity, with $f'(\infty) = 0$.

involved, hence the inclusion of skill parameter s_i. Zero potential cost reduction, labeled as "no R&E potential," yields zero incentive for firms to spend money on R&E, embodied as the assumption $\lim_{\alpha\to 0}\frac{1}{\alpha} f'^{-1}\left(\frac{k}{\alpha}\right)=0$ for $k>0$. The extremes of the cost reduction function could cause optimal R&E spending to have corner solutions, which are ruled out by the assumptions that as r_i approaches zero and infinity, f' approaches infinity and zero respectively. The cost reduction gives firm i an average unit cost of production at time t of $\bar{c}(t)\left[1-f\left(\alpha s_i r_i(t-\delta)\right)\right]$.

Technical know-how may diffuse between firms, and exogenous technological advances may be freely available to all firms, reducing over time the common maximum cost. Hence $\dot{\bar{c}}(t)\le 0$ for all t. Also, $\bar{c}(t)$ is assumed not decline too quickly, so that with a declining industry price $p(t)$, $p(t)-\bar{c}(t)$ is non-increasing. The latter assumption about $\bar{c}(t)$ simplifies the model by ensuring that firms never have an incentive to contract temporarily while waiting for an anticipated future major technological improvement through diffused knowledge.[13]

D. Intentional and Unintentional Exit

Firms may exit the market at any time. The intended time of exit is denoted T_i. ($T_i=\infty$ represents the case of no exit, and a firm that produces 0 output ever after a time τ is defined to exit at $T_i=\tau$.) Exit may also occur for unforeseen reasons such as a plant fire or the death of key personnel; such unintentional exit is termed random exit. Random exit occurs via a Poisson process at (finite) rate $\chi\ge 0$. Thus random exit could occur at any time, with firms having equal chance of random exit to show clearly that differences in χ do not drive the inter-firm differences that will arise from firms' optimal decisions. The probability that a firm has not exited randomly between times $E_i-\delta$ and t, for $t\le T_i$, is denoted $w_i(t)$.

[13] The consequences of an exogenous major technological improvement have been nicely examined elsewhere, for example by Jovanovic and MacDonald (1994). However, such major improvements do not typically appear to be the cause of industry shakeouts, Klepper and Simons (1997, 2000a, 2005). At most, one might expect such technological jumps (which must be anticipable in order to have an effect) to cause only brief departures from implications derived herein.

E. Expected Profit

The expected profit stream of firm i, given the assumptions made above, is

$$\pi_i^e(t) = \left[p(t) - \bar{c}(t)\left[1 - f\left(\alpha s_i r_i(t-\delta)\right)\right]\right] q_i(t) w_i(t) - g\left(\dot{q}_i(t)\right) w_i(t) - e^{\rho\delta} r_i(t-\delta) w_i(t-\delta), \quad (1)$$

where the first term represents expected revenues less production costs, the second term represents expected expansion costs, and the third term represents the present value of the expected R&E spending pertinent to current cost reduction. The third term represents the present value of R&E spending incurred a time δ in the past; this present value is the original amount spent times $e^{\rho\delta}$, where $\rho > 0$ is the discount rate. The sum of these terms motivates firms' choices. Firms choose T_i and the time paths $q_i(t)$ and $r_i(t)$ to maximize the discounted value of the expected profit stream:

$$V_i = \int_{E_i}^{T_i} e^{-\rho t} \pi_i^e(t) dt . \quad (2)$$

Only if $V_i > 0$ does the firm enter.

F. Aggregate Demand

Price is determined by an inverse demand curve, $p(t) = D^{-1}\left(Q(t);t\right)$, where $Q(t)$ is total industry output at time t, with $D^{-1}(Q;t)$ a C^1 function sufficiently high to attract entry (i.e., $V_i > 0$ for some firms at time 0) with a minimum at zero. Market revenue $D^{-1}(Q;t) \cdot Q$ is bounded since buyers have finite funds to spend, price falls with quantity $\frac{\partial D^{-1}(Q;t)}{\partial Q} < 0$, and demand may rise over time $\frac{\partial D^{-1}(Q;t)}{\partial t} \geq 0$, allowing for a diffusion process in which consumers may increasingly desire the product. Demand growth and the random exit probability χ, if sufficiently large, could imply that demand growth outpaces supply; instead, industry-wide results will be shown to hold for sufficiently small values of demand growth and χ, and it is assumed that at all t they accordingly remain limited.

G. Market Equilibrium

A single dominant equilibrium price trend is assumed to be apparent to firms, and this equilibrium is assumed to yield a strictly decreasing time series of price. It will be shown that

such an equilibrium exists. Indeed I conjecture that no equilibrium exists that violates this assumption, and more will be said on this subject later in the paper.

The model applies to a nation's industry facing, at least eventually, international competition. Rising international competition can readily be accommodated in the model, by assuming an amount $Q_C(t)$ of output comes from imports plus in-country production facilities. The international competitors' output $Q_C(t) \geq 0$ is assumed to be nondecreasing, continuous with t, and sufficiently small that some in-country production remains (at least until the patterns described in the paper have occurred). After some time κ arbitrarily far in the future, expansion by international competitors is sufficient to ensure $\dot{p} < 0$. The latter assumption guarantees existence of the market equilibrium, although the assumption may not be crucial as the above conjecture implies. Hence the results hold for nations open, initially or eventually, to rising worldwide competition.

H. Atomism

The model has assumed that all firms are atomistic, so that no firm influences either the price $p(t)$ or the maximal cost $\bar{c}(t)$. This assumption deliberately abstracts from oligopolistic behavior in order to prove that resulting patterns of concentration can be driven by pure forces of competition. The results thus provide a baseline against which oligopolistic behavior can be compared, and a building block for future models. A later section of the paper returns briefly to assess partial departures from this baseline in eventual situations of oligopoly.

III. Firm and Industry Dynamics[14]

This section of the paper proves implications of the model, through a series of lemmas and theorems organized for summary or in-depth reading. For key conclusions of a subsection, read the first paragraphs of the subsection. For more detail read the assertions of theorems or, if desired, the full proofs.

A. Optimal Firm Decisions

Profit-maximizing behavior by firms yields, using optimal control theory, optimal firm decisions. The optimal decisions are catalogued in Theorem 1, which has four parts. Part A

[14] All analytic calculations in this paper have been double-checked using Mathematica and Maple. The model has been checked to ensure dimensional consistency.

establishes the optimal decision for R&E spending. Part B establishes the optimal time path of expansion. Part C establishes the optimal exit time. Part D establishes another way to represent the optimal expansion conditions, using the costate variable of optimal control theory. Existence and uniqueness is established for the R&E spending decision, uniqueness is proven for the expansion decision, and later, in Theorem 2 of the next subsection, uniqueness will be shown for the exit decision. This subsection thus lays out in the paper's first theorem all the requirements of optimal behavior.

In each part of Theorem 1, the optimal decision involves balancing marginal benefits against marginal costs. Optimal R&E spending occurs when the marginal cost saving from R&E just equals the present value of an extra dollar per year spent (δ years ago) on R&E. Optimal expansion occurs when the marginal increase in the value accumulated from t to T_i just equals the marginal cost of expansion. Optimal exit occurs when the firm's cost flow just equals its revenue stream.

In optimal control theory, functions $\lambda_i(t)$ and $\eta_i(t)$ arise. $\lambda_i(t) = \dfrac{\partial V_i}{\partial q_i(t)}$ is known as the costate variable. It gives the value associated with an increase in the state $q_i(t)$ at time t, assuming the state is adjusted optimally thereafter. The similar function $\eta_i(t)$ is associated with the constraint $q_i(t) \geq 0$ and can be nonzero only when $q_i(t) = 0$. $\lambda_i(t)$ and $\eta_i(t)$ are used throughout the proofs.

Two more terms used throughout the proofs describe effective values of the discount rate and profit. Both values are "effective" in that they incorporate effects of random exit. The effective discount rate is $\psi = \rho + \chi$. The effective profit, defined in (3), is written $\pi_i(t)$ and turns out to be $\pi_i(t) = e^{\chi t}\pi_i^e(t)$. The effective profit should be discounted at rate ψ, since integrating $e^{-\psi t}\pi_i(t)$ yields the firm's expected value V_i from (2). These properties are shown in Lemma 1.

Lemma 1. Define $\psi = \rho + \chi$ and

$$\pi_i(t) = e^{\chi(E_i-\delta)}\left(\left[\, p(t) - \overline{c}(t)\left[1 - f\left(\alpha s_i r_i(t-\delta)\right)\right]\right] q_i(t) - g\left(\dot{q}_i(t)\right) - e^{\psi\delta} r_i(t-\delta)\right). \qquad (3)$$

Firm i's effective profit is related to its expected profit according to:

$$\pi_i(t) = e^{\chi t}\pi_i^e(t)\,. \tag{4}$$

Firm i's cumulative discounted expected profit equals

$$V_i = \int_{E_i}^{T_i} e^{-\psi t}\pi_i(t)dt\,. \tag{5}$$

Proof: The Poisson random exit process can be written

$$\frac{dw_i(t)}{dt} = -\chi w_i(t)\,. \tag{6}$$

The probability $w_i(t)$ starts at 1 at time $E_i - \delta$: $w_i(E_i - \delta) = 1$. Rewrite (6) as $\frac{dw_i(t)}{w_i(t)} = -\chi dt$, integrate both sides, and choose a constant of integration that yields the initial probability of 1, and the resulting solution to the differential equation (6) is

$$w_i(t) = e^{-\chi(t-(E_i-\delta))}\,. \tag{7}$$

This finding is equivalent to a basic result in statistical survival analysis. Substitute the survival probability (7) into (1) and multiply both sides by $e^{\chi t}$ to obtain the rewritten profit function (4). Substitute (4) into (2) to obtain the rewritten value function (5). ■

Theorem 1. Firm i's optimal decisions are as follows:[15]

A. Firm i's time path of optimal R&E spending $r_i^*(t-\delta)$ satisfies

$$f'\left(\alpha s_i r_i^*(t-\delta)\right) = \frac{e^{\psi\delta}}{\alpha s_i \bar{c}(t) q_i(t)} \tag{8}$$

for each time t when the firm is a producer (at other times, $r_i^*(t-\delta) = 0$). $r_i^*(t-\delta)$ exists and is unique for all t. If $p(t)$ is continuous, $r_i^*(t-\delta)$ is continuous.

B. Firm i's optimal time path of expansion $\dot{q}_i^*(t)$, following its entry at time E_i with output $q_i(E_i) = 0$, satisfies

[15] Parts B through D of Theorem 1 also hold for any exogenous (i.e., independent of $q_i(t)$ and $\dot{q}_i(t)$) time path of R&E spending, by substituting that time path in place of the optimal time path $r_i^*(t-\delta)$.

$$\ddot{q}_i^*(t) + \frac{p(t) - \overline{c}(t)\left[1 - f\left(\alpha s_i r_i^*(t-\delta)\right)\right] - \psi g'\left(\dot{q}_i^*(t)\right) + e^{-\chi(E_i-\delta)+\psi t}\eta_i(t)}{g''\left(\dot{q}_i^*(t)\right)} = 0, \tag{9}$$

where $\eta_i(t) \geq 0$ and $\eta_i(t)\, q_i(t) = 0$, subject to the boundary condition

$$\dot{q}_i^*(T_i^*) = 0. \tag{10}$$

If $p(t)$ is continuous, the time paths $q_i^*(t)$, $\dot{q}_i^*(t)$, $\ddot{q}_i^*(t)$, and $\pi_i(t)$ are unique and continuous whenever $q_i^*(t) > 0$ (and at any τ when $q_i^*(\tau) = 0$ and $q_i^*(t) > 0$ in the rest of a neighborhood around τ), with $q_i^*(t)$ always continuous.

C. Firm i's optimal exit time T_i^* occurs when

$$\pi_i(T_i^*) = 0. \tag{11}$$

D. The Euler equation (9) is equivalent to the two first-order differential equations:

$$g'\left(\dot{q}_i^*(t)\right) = e^{\psi t - \chi(E_i-\delta)}\lambda_i(t), \tag{12}$$

$$\dot{\lambda}_i(t) = -e^{-\psi t + \chi(E_i-\delta)}\left(p(t) - \overline{c}(t)\left[1 - f\left(\alpha s_i r_i^*(t-\delta)\right)\right]\right) - \eta_i(t). \tag{13}$$

Also,

$$\lambda_i(T_i^*) = 0. \tag{14}$$

If $p(t)$ is continuous, $\lambda_1(t)$ is continuous whenever $q_i^*(t) > 0$ (and at any τ when $q_i^*(\tau) = 0$ and $q_i^*(t) > 0$ in the rest of a neighborhood around τ).

(The above holds regardless whether or not $p(t)$ is strictly decreasing.)

Proof: The proof is a straightforward application of optimal control theory. It involves equations between control variables, state variable, terminal time, and Hamiltonian. The control variables are $r_i(t-\delta)$ and $\dot{q}_i(t)$. The terminal time T_i is at the firm's discretion, and the terminal state $q_i(T_i)$ is constrained only to be nonnegative. For the time rate of change of the state variable, $q_i(t)$, in terms of the controls, write $\zeta\left(r_i(t-\delta), \dot{q}_i(t)\right)$; in this model $\zeta\left(r_i(t-\delta), \dot{q}_i(t)\right) = \dot{q}_i(t)$.

The Hamiltonian is defined as $\mathcal{H}_i(t) = \lambda_{i0}\left[e^{-\psi t}\pi_i(t) + \lambda_i(t)\zeta\left(r_i(t-\delta), \dot{q}_i(t)\right)\right] + \eta_i(t)q_i(t)$.[16] The necessary conditions from optimal control theory are:[17]

$$\frac{\partial \mathcal{H}_i(t)}{\partial q_i(t)} = -\dot{\lambda}_i(t), \tag{15}$$

$$\frac{\partial \mathcal{H}_i(t)}{\partial r_i(t-\delta)} = 0, \tag{16}$$

$$\frac{\partial \mathcal{H}_i(t)}{\partial \dot{q}_i(t)} = 0, \tag{17}$$

$$\mathcal{H}_i(t) \text{ is maximized by } r_i(t-\delta) \text{ and } \dot{q}_i(t), \tag{18}$$

$$\lambda_{i0}\,\lambda_i(t) \neq 0, \tag{19}$$

$$\lambda_{i0} = 0 \text{ or } 1, \tag{20}$$

$$\lambda_i(t) \geq 0, \tag{21}$$

$$\eta_i(t)\,q_i(t) = 0, \tag{22}$$

$$\eta_i(t) \geq 0, \tag{23}$$

$$\lambda_i(T_i) = 0, \text{ and} \tag{24}$$

$$\mathcal{H}_i(T_i) = 0. \tag{25}$$

The necessary conditions (15) through (25) imply results (8) through (14). For $\lambda_{i0}\,\lambda_i(T_i)$ to satisfy (19), (24) implies $\lambda_{i0} \neq 0$ and hence (20) implies $\lambda_{i0} = 1$. Compute the derivative in (15), (16), or (17) respectively and solve for $\dot{\lambda}_i(t)$, $f'\left(\alpha s_i r_i(t-\delta)\right)$, or $\dot{q}_i(t)$ to obtain (13), (8), or (12). Solve (12) for $\lambda_i(t)$, compute the total derivative with respect to t, substitute into (13) to eliminate the variable $\lambda_i(t)$, and rearrange to obtain (9). Equation (24) is identical to (14). Let $t = T_i^*$ in (12) and substitute for $\lambda_i(T_i^*)$ using (14) to obtain (10). Substitute (14) into (25) and

[16] These conditions apply on the domain of t for firm i, $t \in [E_i, T_i^*]$.

[17] See for example Seierstad and Sydsæter (1987) and Kamien and Schwartz (1991, especially pp. 160, 219, and 231).

apply (22) to obtain (11). The maximization condition (18) is satisfied, since $\frac{\partial^2 \mathcal{H}_i(t)}{\partial r_i(t-\delta)^2} < 0$, $\frac{\partial^2 \mathcal{H}_i(t)}{\partial \dot{q}_i(t)^2} < 0$, and $\frac{\partial^2 \mathcal{H}_i(t)}{\partial \dot{q}_i(t) \partial r_i(t-\delta)} = 0$ so that $\mathcal{H}_i(t)$ is concave in $r_i(t-\delta)$ and $\dot{q}_i(t)$.

It remains to prove, using facts about differential equations, the existence, uniqueness, and continuity claims. Uniqueness and continuity of the firm's optimal expansion time path $q_i^*(t)$ follows from the fact that for any n^{th} order differential equation $\frac{d^n q}{dt^n} = h(t, q, \frac{dq}{dt}, ..., \frac{d^{n-1} q}{dt^{n-1}})$, with $h(\cdot)$ continuous and continuously differentiable in its last n arguments, any solution has unique continuous time paths for $q(t)$ and its first $n-1$ derivatives (cf., Kamien and Schwartz, 1991, p. 351). Uniqueness and continuity of $\lambda_i(t)$ and $\dot{q}_i^*(t)$ are guaranteed by a similar theorem for systems of differential equations. Continuity of $\lambda_i(t)$ could be broken at times when $q_i^*(t)$ transitions to or from 0 and remains at 0 for a positive length of time, as optimal control theory indicates given state variable inequality constraints, hence the guarantee of continuity only when $q_i^*(t) > 0$ and, if $p(t)$ is continuous, at any τ when $q_i^*(\tau) = 0$ and $q_i^*(t) > 0$ immediately before and after τ (continuity can actually be proven more generally, but for the case mentioned see Seierstad and Sydsæter's (1987) comment on state constraints with continuous derivatives, bottom of p. 318, and apply their conditions to ensure continuity in note 3e, p. 334). Existence and uniqueness of $r_i^*(t-\delta)$ follows from (8) and from the assumptions about $f(\cdot)$. Since $f'(\cdot)$ decreases continuously in its argument, with its domain the entire set of positive numbers as well as positive infinity, a unique value $r_i^*(t-\delta)$ exists that satisfies equation (8) for every possible value of the equation's right-hand side. Hence $r_i^*(t-\delta)$ exists and is unique and continuous. Uniqueness and continuity of $\ddot{q}_i^*(t)$ and $\pi_i(t)$ follows given the continuity of all other terms in (9) and (3) respectively. ■

B. Expansion at All Times Before Exit

Having established the optimal decisions of firms, a next and much more difficult task is to discern what these decisions mean in terms of firm growth. Most implications will be proven to hold if $p(t)$ is continuous and $\dot{p}(t) < 0$ for all t, preconditions later shown to be true.

Theorem 2 proves that $\lambda_i(t) > 0$ while the firm produces and, in consequence, that each firm expands continually until it exits. Three lemmas first establish intermediate results.

Some lemmas and theorems use the abbreviation $f^*(t) = f\left(\alpha s_i r_i^*(t-\delta)\right)$ for the optimal fractional cost reduction.

Lemma 2. $\lambda_i(t)$ has the same sign as $\dot{q}_i^*(t)$, and $\dfrac{\partial \dot{q}_i^*(t)}{\partial \lambda_i(t)} > 0$.

Proof: Equation (17) is $\lambda_i(t) = e^{-\psi t + \chi(E_i - \delta)} g'\left(\dot{q}_i^*(t)\right)$. The assumptions about $g(\cdot)$ imply $g'\left(\dot{q}_i^*(t)\right)$ is negative for $\dot{q}_i^*(t) < 0$, zero for $\dot{q}_i^*(t) = 0$, positive for $\dot{q}_i^*(t) > 0$, and strictly increasing everywhere. Therefore a one-to-one mapping exists between $\dot{q}_i^*(t)$ and $\lambda_i(t)$, with $\dot{q}_i^*(t)$ having the same sign as $\lambda_i(t)$ and $\dfrac{\partial \dot{q}_i^*(t)}{\partial \lambda_i(t)} > 0$. ■

Lemma 3. Suppose that $\dot{p}(t) < 0$ at some t. $\dot{p}(t) - \dot{\bar{c}}(t) + \dot{\bar{c}}(t) f(\alpha s_i r_i(t-\delta)) < 0$, if $r_i(t-\delta) > 0$ (which is guaranteed if $r_i(t-\delta) = r_i^*(t-\delta)$ and $q_i(t) > 0$).

Proof: $p(t) - \bar{c}(t)$ and $\bar{c}(t)$ are non-increasing, implying $\dot{p}(t) - \dot{\bar{c}}(t) \le 0$ and $\dot{\bar{c}}(t) \le 0$. There is no possibility that both $\dot{p}(t) - \dot{\bar{c}}(t) = 0$ and $\dot{\bar{c}}(t) = 0$ since $\dot{p}(t) < 0$. Since $r_i(t-\delta) > 0$ guarantees $f(\alpha s_i r_i(t-\delta)) > 0$, the Lemma follows directly. ■

Lemma 4. Suppose that $\dot{p}(t) < 0$ for all t from t_1 to $t_2 + D$, except $\dot{p}(t_1) \le 0$. Compare (a) the value $W_i(0)$ obtained by producing with $q_i(t)$, $\dot{q}_i(t)$, and $r_i(t-\delta)$ from $t = t_1$ to t_2 versus (b) the value $W_i(D)$ obtained by following the same paths delayed by D time units, $q_i(t-D)$, $\dot{q}_i(t-D)$, and $r_i(t-D-\delta)$ from $t = t_1 + D$ to $t_2 + D$.

A. $W_i'(D) < 0$ at any point $D \geq 0$ for which $W_i(D) \geq 0$.

B. If $D > 0$ and $W_i(D) \geq 0$ then $W_i(0) > W_i(D)$.

Proof: First it is shown that, holding firm-specific traits and decisions constant, effective profit decreases with time. Let $\tilde{\pi}_i(t, D)$ represent the firm's effective profit function at time $t + D$ but using $q_i(t)$, $\dot{q}_i(t)$, and $r_i(t - \delta)$ instead of $q_i(t + D)$, $\dot{q}_i(t + D)$, and $r_i(t + D - \delta)$. Compare this profit function at times $t + D$ as D is varied, but t is held constant. Since t is constant, the output, growth rate, and R&E expenditure remain constant. Differentiating yields

$$\begin{aligned} \frac{\partial \tilde{\pi}_i(t, D)}{\partial D} &= e^{\chi(E_i - \delta)} \left(\dot{p}(t + D) - \dot{\bar{c}}(t + D)\left[1 - f\left(\alpha s_i r_i(t - \delta)\right)\right]\right) q_i(t) \\ &= e^{\chi(E_i - \delta)} \left(\dot{p}(t + D) - \dot{\bar{c}}(t + D) + \dot{\bar{c}}(t + D) f\left(\alpha s_i r_i(t - \delta)\right)\right) q_i(t) < 0. \end{aligned} \tag{26}$$

Expression (26) is negative by Lemma 3.

Next cumulative discounted profit, i.e., value, is shown to decrease with time. $W_i(D)$ is the value accumulated from $t_1 + D$ to $t_2 + D$ using the delayed policies: $W_i(D) = \int_{t_1}^{t_2} e^{-\psi(t+D)} \tilde{\pi}_i(t, D) dt$. As the start time $t_1 + D$ is changed, the value changes according to

$$\begin{aligned} \frac{\partial W_i(D)}{\partial D} &= \frac{\partial}{\partial D} \int_{t_1}^{t_2} e^{-\psi(t+D)} \tilde{\pi}_i(t, D) dt = \int_{t_1}^{t_2} \frac{\partial}{\partial D} \left(e^{-\psi(t+D)} \tilde{\pi}_i(t, D)\right) dt \\ &= -\psi \int_{t_1}^{t_2} e^{-\psi(t+D)} \tilde{\pi}_i(t, D) dt + \int_{t_1}^{t_2} \left(e^{-\psi(t+D)} \frac{\partial \tilde{\pi}_i(t, D)}{\partial D}\right) dt. \end{aligned} \tag{27}$$

In the last line of (27), the first term is $-\psi W_i(D)$, and is therefore non-positive if $W_i(D) \geq 0$, or strictly negative if $W_i(D) > 0$. The second term is an integral of the discounted partial derivative $\frac{\partial \tilde{\pi}_i(t, D)}{\partial D}$, and is strictly negative because (26) implies $\frac{\partial \tilde{\pi}_i(t, D)}{\partial D} < 0$ for all $t > t_1$ (with $\frac{\partial \tilde{\pi}_i(t_1, D)}{\partial D} \leq 0$). Hence, combining the two terms, $W_i(D) \geq 0$ implies $W_i'(D) < 0$. This proves part A of the Lemma.

To prove part B of the Lemma, use part A and note that $W_i(D) - W_i(0) = \int_0^D W_i'(\Delta) d\Delta$. Start at $\Delta = D > 0$ with $W_i(D) \geq 0$, and decrease Δ from D to 0. $W_i(\Delta)$ increases initially as

Δ is decreased from D, using part A, since $W_i'(D) < 0$. By recursion $W_i(\Delta)$ continually increases thereafter as Δ is decreased to 0, still using part A, since $W_i'(\Delta) < 0$ throughout. With $W_i'(\Delta) < 0$ for all Δ from 0 to D, $W_i(D) - W_i(0) = \int_0^D W_i'(\Delta) d\Delta < 0$. Add $W_i(0)$ to the left- and right-hand sides of this inequality to obtain $W_i(0) > W_i(D)$. ■

Lemma 5. Suppose that $\dot{p}(t) < 0$. If $\dot{q}_i(t) = 0$ and $q_i(t) > 0$, then $\dot{\pi}_i(t) < 0$.

Proof: Use (3), with $r_i(t-\delta) = r_i^*(t-\delta)$, totally differentiate with respect to t, and replace $f'\left(\alpha s_i r_i^*(t-\delta)\right)$ using (8) with the result that two terms cancel, to obtain:

$$\frac{d\pi_i(t)}{dt} = e^{\chi(E_i-\delta)}\left(\left(p(t) - \bar{c}(t)\left[1 - f^*(t)\right]\right)\dot{q}_i(t) + \left(\dot{p}(t) - \dot{\bar{c}}(t)\left[1 - f^*(t)\right]\right)q_i(t) - g'\left(\dot{q}_i(t)\right)\ddot{q}_i(t)\right). \quad (28)$$

Substitute $\dot{q}_i(t) = 0$ to obtain:

$$\frac{d\pi_i(t)}{dt} = e^{\chi(E_i-\delta)}\left(\dot{p}(t) - \dot{\bar{c}}(t) + \dot{\bar{c}}(t) f^*(t)\right)q_i(t). \quad (29)$$

Expression (29) is, using Lemma 3, strictly negative. Therefore if $\dot{q}_i(t) = 0$, $\dot{\pi}_i(t) < 0$. ■

Lemma 6. Suppose that $\dot{p}(t) < 0$ and p(t) is continuous. The following results hold:

A. For all t from E_i to T_i^*, either $\pi_i(t) \geq 0$ or $\dot{q}_i^*(t) > 0$ (or both).

B. $\pi_i(t) = 0$ and $\dot{q}_i^*(t) = 0$ do not simultaneously occur throughout an interval $(\tau, T_i^*]$, for any $\tau < T_i^*$.

Proof: Part A of the Lemma is proved in five steps, each pertaining to a situation when, for some t, $\pi_i(t) < 0$ and $\dot{q}_i^*(t) \leq 0$. The five steps are illustrated in Figure 1: 1. a time interval throughout which $\pi_i(t) \leq 0$ cannot exist at the end of the firm's production history, 2. a flat or valley-shaped region of $q_i^*(t)$ cannot coincide with an interval throughout which $\pi_i(t) \leq 0$, 3. a region ending at the bottom of a valley in $q_i^*(t)$ cannot coincide with an interval throughout

which $\pi_i(t) \le 0$, 4. a region on the left side of a valley in $q_i^*(t)$ cannot coincide with an interval throughout which $\pi_i(t) \le 0$, and 5. no single point in time preceding T_i^* has both $\pi_i(t) < 0$ and $\dot{q}_i^*(t) \le 0$. Part B of the Lemma is established during the first step.

Step 1. Suppose there is a time period (τ, T_i^*) throughout which $\pi_i(t) \le 0$. If at any time during this period $\pi_i(t) < 0$, then the firm could increase its value V_i by exiting earlier, so since T_i^* is optimal it is impossible that $\pi_i(t) < 0$ at any time during the period. This leaves only the possibility $\pi_i(t) = 0$ throughout the period. During the period, at any time t preceding T_i^*, $\dot{q}_i^*(t) \neq 0$ or $q_i^*(t) = 0$. To see this, note that if $\dot{q}_i^*(t) = 0$ and $q_i^*(t) \neq 0$ at any t then, by Lemma 5, $\dot{\pi}_i(t) < 0$, yielding at a later time during the period negative profit, which has just been shown to be impossible. The only remaining possibilities are (i) $\dot{q}_i^*(t) < 0$ throughout the period, (ii) $\dot{q}_i^*(t) > 0$ throughout the period, or (iii) $q_i^*(t) = 0$ at some time during the period and thereafter $\dot{q}_i^*(t) > 0$ for a time interval ending at T_i^*, because by Theorem 1B $\dot{q}_i^*(t)$ is continuous and hence cannot cross the point $\dot{q}_i^*(t) = 0$ unless $q_i^*(t) = 0$ (the definition of T_i^* rules out (iv) $q_i^*(t) = 0$ for a time interval ending at T_i^*). But whenever $\dot{q}_i^*(t) < 0$ or $\dot{q}_i^*(t) > 0$, the firm could instead choose $\dot{q}_i(t) = 0$, improving profit by yielding $\pi_i(t) = g\left(q_i^*(t)\right)$ instead of $\pi_i(t) = 0$, with positive profit persisting for some time thereafter (given continuity), allowing the firm to increase its value V_i through appropriate choice of T_i and implying that $\dot{q}_i^*(t)$ is not optimal, a contradiction. Hence there cannot be a time period at the end of the firm's history in which $\pi_i(t) \le 0$, except for the exit time T_i^*.

Step 2. Suppose the firm's production history contains a period throughout which $\pi_i(t) \le 0$ and $q_i^*(t)$ is flat or valley-shaped, i.e., $\dot{q}_i^*(t) \le 0$ (for a positive length of time) initially during the period, then $\dot{q}_i^*(t) \ge 0$ (for a positive length of time) during the remainder of the period. Denote the times when this period begins and ends as a and b respectively, let $\tilde{q} = \min\left(q_i^*(a), q_i^*(b)\right)$, and denote the first and last times during the period when $q_i^*(t) = \tilde{q}$ as c and d respectively. The firm could increase its value V_i by, as shown in Lemma 4B, choosing a

new curve $q_i(t) = q_i^*(t + D)$, where $D = d - c$, beginning at time c, and by exiting at $T_i^* - D$. This implies $q_i^*(t)$ is not optimal, which is a contradiction, so a period with $\pi_i(t) \le 0$ during which $q_i^*(t)$ is flat or valley-shaped never arises in the firm's production history.

Step 3. Suppose the firm's production history contains a period throughout which $\pi_i(t) \le 0$ and $\dot{q}_i^*(t) \le 0$, and that immediately after this period $\dot{q}_i^*(t) > 0$. Lemma 2 implies that during the period $\lambda_i(t) \le 0$, and immediately afterward $\lambda_i(t) > 0$. Hence $\dot{\lambda}_i(t) > 0$ at the end of the interval, and by (13) in Theorem 1D, $p(t) - \overline{c}(t)\left[1 - f\left(\alpha s_i r_i^*(t - \delta)\right)\right] < 0$. From step 2 above, immediately after the period it cannot be the case that $\pi_i(t) \le 0$, so $\pi_i(t) > 0$ immediately after the period. But since $\pi_i(t)$ is continuous (given the continuity of $q_i^*(t)$ and $\dot{q}_i^*(t)$ from Theorem 1B and the continuity of $p(t)$), it follows that at one or more points in time at the end of the interval, $\pi_i(t) = 0$, which is impossible given that $p(t) - \overline{c}(t)\left[1 - f\left(\alpha s_i r_i^*(t - \delta)\right)\right] < 0$. This contradiction shows that a period with $\pi_i(t) \le 0$ during which $q_i^*(t)$ is at the bottom left side of a valley never arises in the firm's production history.

Step 4. Suppose the firm's production history contains a period throughout which $\pi_i(t) \le 0$ and $\dot{q}_i^*(t) \le 0$, and that immediately after this period $\pi_i(t) > 0$. From step 3 above, immediately after the period it cannot be the case that $\dot{q}_i^*(t) > 0$, so $\dot{q}_i^*(t) \le 0$ immediately after the period. Let τ denote the time when the period ends. At the end of the period there will be shown to exist an interval $[a, b]$ with two properties that together allow the firm to improve its profit, contradicting the optimality of $\dot{q}_i^*(t)$.

The first property that holds in the interval $[a, b]$ is $\left[p(t) - \overline{c}(t)\left[1 - f\left(\alpha s_i r_i^*(t - \delta)\right)\right]\right] q_i^*(t) - e^{\psi\delta} r_i^*(t - \delta) > 0$. This property holds at $t = \tau$, using (3) while noting that $g\left(\dot{q}_i^*(\tau)\right) > 0$ and $\pi_i(\tau) = 0$. Moreover, there is an interval $[a_1, b_1]$, such that $a_1 < \tau < b_1$, throughout which the same inequality holds, since $\dot{q}_i^*(t)$, $r_i^*(t - \delta)$, and other constituents of $\pi_i(t)$ are all continuous. Choose any such a_1 and b_1.

The second property that holds in the interval $[a,b]$ is $\ddot{q}_i^*(t) > 0$. This will be shown using properties about $\pi_i(t)$, $\dot{\pi}_i(t)$, $q_i^*(t)$, and $\dot{q}_i^*(t)$ around $t = \tau$. $\pi_i(\tau) = 0$ and $\dot{\pi}_i(\tau) > 0$, since $\pi_i(t)$ is continuous and becomes positive immediately after τ. The rate of change in profit at τ is, using (28),

$$\dot{\pi}_i(\tau) = e^{\chi(E_i-\delta)}\left(\left(p(\tau)-\overline{c}(\tau)\left[1-f^*(\tau)\right]\right)\dot{q}_i^*(\tau)+\left(\dot{p}(\tau)-\dot{\overline{c}}(\tau)\left[1-f^*(\tau)\right]\right)q_i^*(\tau)-g'\left(\dot{q}_i^*(\tau)\right)\ddot{q}_i^*(\tau)\right).$$

Note that $q_i^*(\tau) > 0$, because if $q_i^*(\tau) = 0$ the firm could not achieve positive profit immediately after τ since $\dot{q}_i^*(t) \le 0$ around $t = \tau$. Therefore $p(\tau)-\overline{c}(\tau)\left[1-f^*(\tau)\right] > 0$, since using (3) this is the only way to achieve $\pi_i(\tau) = 0$ when $q_i^*(\tau) > 0$. Hence, given $\dot{q}_i^*(\tau) \le 0$, the first term in $\dot{\pi}_i(\tau)$ is nonpositive. The second term in $\dot{\pi}_i(\tau)$ is strictly negative, using Lemma 3. Therefore $\dot{\pi}_i(\tau) > 0$ requires $-g'\left(\dot{q}_i^*(\tau)\right)\ddot{q}_i^*(\tau) > 0$. In this expression, $g'\left(\dot{q}_i^*(\tau)\right) \le 0$ since $\dot{q}_i^*(\tau) \le 0$, and hence $\dot{\pi}_i(\tau) > 0$ requires $\ddot{q}_i^*(\tau) > 0$. Using (9), $\ddot{q}_i^*(t)$ is continuous and hence there is an interval $[a_2,b_2]$, such that $a_2 < \tau < b_2$, throughout which $\ddot{q}_i^*(t) > 0$. Choose any such a_2 and b_2.

Define b_3 to be the first time after τ such that $\dot{q}_i^*(b_3) \ge 0$ ($b_3 > \tau$ since $\dot{q}_i^*(\tau)$ cannot be 0 immediately after τ using step 2 above). Define $a = \max(a_1,a_2)$ and $b = \min(b_1,b_2,b_3)$, so that $\dot{q}_i^*(t) < 0$, and the two properties hold, throughout $[a,b]$. Divide the interval in half, forming subintervals $[a,a+\frac{b-a}{2}]$ and $[a+\frac{b-a}{2},b]$. The firm can enhance its cumulative discounted profit by maintaining $r_i(t-\delta) = r_i^*(t-\delta)$ for all t in $[a,b]$ but swapping its optimal growth rates in the two subintervals, that is, choosing $\dot{q}_i(t) = \dot{q}_i^*\left(t+\frac{b-a}{2}\right)$ during the first subinterval and choosing $\dot{q}_i(t) = \dot{q}_i^*\left(t-\frac{b-a}{2}\right)$ during the second subinterval. Figure 2 illustrates the growth rates resulting from this policy. To see that this enhances the firm's value, substitute (3) into (5) and note (i) the new policy increases $\int_a^b e^{-\psi t}e^{\chi(E_i-\delta)}\left[p(t)-\overline{c}(t)\left[1-f\left(\alpha s_i r_i(t-\delta)\right)\right]\right]q_i(t)dt$ because it implies greater firm sizes at all t in $[a,b]$, during which the price-cost margin is strictly positive, (ii) the new policy increases

$\int_a^b e^{-\psi t} e^{\chi(E_i-\delta)} \left(-g\left(\dot{q}_i(t)\right)\right) dt$ because it moves the more negative growth rates, with their larger values of $g\left(\dot{q}_i(t)\right)$, later in time when they are discounted more, (iii) the new policy does not affect $\int_a^b e^{-\psi t} e^{\chi(E_i-\delta)} \left(-e^{\psi\delta} r_i(t-\delta)\right) dt$ because it does not affect the choice of $r_i(t-\delta)$, and (iv) the new policy does not affect profit before time a nor after time b because it does not affect $q_i(t)$, $\dot{q}_i(t)$, or $r_i(t-\delta)$ for $t < a$ or $t > b$. Since the firm can improve its value by choosing a path $\dot{q}_i(t) \neq \dot{q}_i^*(t)$, $\dot{q}_i^*(t)$ cannot be optimal. This contradiction rules out the case in Figure 1 involving intervals with $\pi_i(t) \leq 0$ on the left-hand side of a valley in $q_i^*(t)$.

Step 5. Suppose there is a time τ when $\pi_i(\tau) < 0$ and $\dot{q}_i^*(\tau) \leq 0$. It will be shown that this is impossible, because at the margin, the firm can enhance its value by choosing an alternative policy for $\dot{q}_i(t)$ and T_i. Steps 1-4 above have shown that there exists no positive-length interval throughout which $\pi_i(t) \leq 0$ and $\dot{q}_i^*(t) \leq 0$, so given the continuity of $\dot{q}_i^*(t)$, it follows that $\dot{q}_i^*(\tau) = 0$. By continuity there exists an interval $[a,b]$ around τ $(a < \tau < b)$ such that $\pi_i(t) < 0$ throughout $[a,b]$, $\dot{q}_i^*(t) > 0$ throughout $[a,\tau)$ and $(\tau,b]$, $\ddot{q}_i^*(t) < 0$ for $t \in [a,\tau)$, $\ddot{q}_i^*(t) > 0$ for $t \in (\tau,b]$, and $\dot{\lambda}(t) < 0$ throughout $[a,\tau)$. Within the interval, since $\lambda(t)$ has the same sign as $\dot{q}_i^*(t)$ by Lemma 2, $\lambda(t) > 0$ before and after τ with $\lambda(\tau) = 0$, and hence by continuity $\dot{\lambda}(t) < 0$ immediately before τ. The alternative strategy is to choose $\dot{q}_i(t) = \dot{q}_i^*(t) + \frac{1}{\Delta}\int_\tau^{\tau+\Delta} \dot{q}_i^*(t)\, dt$ throughout the subinterval $[\tau-\Delta,\tau)$, for an appropriate $\Delta > 0$, thus yielding $q_i(\tau) = q_i^*(\tau+\Delta)$, to choose $r_i(t) = r_i^*(t)$ through the same subinterval, to choose $q_i(t) = q_i^*(t+\Delta)$ and $r_i(t) = r_i^*(t+\Delta)$ thereafter from τ to $T_i^* - \Delta$, and then to exit at $T_i = T_i^* - \Delta$.

Consider the resulting change in value over the firm's production history, and in the resulting expression examine each of the terms. The change in value, relative to the optimal value, is

$$W_i(0) - W_i(\Delta) - \int_{\tau}^{\tau+\Delta} e^{-\psi t}\pi_i(t)dt$$
$$+ \int_{\tau-\Delta}^{\tau} e^{-\psi t}\left[\, p(t) - \overline{c}(t)\left[1 - f^*(t)\right]\right]\left(q_i(t) - q_i^*(t)\right)dt$$
$$- \Delta g\left(\dot{q}_i^*(\tau-\Delta)\right)\frac{e^{-\psi(\tau-\Delta)} - e^{-\psi\tau}}{\psi} + \int_{\tau-\Delta}^{\tau} e^{-\psi t} g(q_i^*(t))dt\,, \qquad (30)$$

where $W_i(\Delta)$ is the value accumulated by pursuing the optimal policies from $\tau+\Delta$ to T_i^* and $W_i(0)$ is the value accumulated by pursuing the same policies Δ units of time earlier, from τ to $T_i^* - \Delta$. The first three terms in (30) represent the change in value from τ onward, whereas the last three terms represent the change in value before τ. The first term is the value accumulated from τ onward under the alternative policy, and from this the second and third terms subtract the value accumulated from τ onward under the optimal policy. The sum of the first two terms, $W_i(0) - W_i(\Delta)$, is strictly positive by Lemma 4B. The third term is nonnegative since $\pi_i(t) \le 0$ throughout the interval. The fourth term, which stems from the faster growth and hence larger size under the alternative policy during the subinterval $[\tau-\Delta, \tau)$, represents the difference in (discounted) accumulated revenue and production cost between the two policies. In the fourth term $q_i(t) - q_i^*(t) > 0$ (or $= 0$ at $t = \tau - \Delta$) and $p(t) - \overline{c}(t)\left[1 - f^*(t)\right] > 0$ by (13) since $\dot{\lambda}(t) < 0$ and $\eta_i(t) = 0$, so the fourth term is strictly positive. The fifth and sixth terms represent a difference in (discounted) expenditures on growth under the alternative versus optimal policy during the subinterval $[\tau-\Delta, \tau)$, with the fifth term corresponding to the alternative policy and the sixth term corresponding to the optimal policy. The fifth term is negative and the sixth term is positive. The growth costs of the alternative policy exceed those of the optimal policy, so the sum of the fifth and sixth terms is negative. However, it will be shown that for Δ sufficiently small, the sum of the first two terms (positive) exceeds the absolute value of the fifth term and hence (30) is strictly positive.

To see that the first two terms dominate the fifth for Δ sufficiently small, compute the derivative of each expression with respect to Δ, compare the resulting derivatives, and apply a continuity argument. The sum of the first two terms, $W_i(0) - W_i(\Delta)$, has derivative

$$\frac{d}{d\Delta}\left(W_i(0) - W_i(\Delta)\right) = -W_i'(\Delta)\,, \qquad (31)$$

which is strictly positive at $\Delta = 0$ using Lemma 4A. Writing the fifth term as $h(\Delta)$, the fifth term has derivative

$$h'(\Delta) = \frac{e^{-\psi\tau}}{\psi}\left(\Delta g'\left(\dot{q}_i^*(\tau-\Delta)\right)\ddot{q}_i^*(\tau-\Delta)[e^{\psi\Delta}-1] - g\left(\dot{q}_i^*(\tau-\Delta)\right)\left[e^{\psi\Delta}(1+\psi\Delta)-1\right]\right), \qquad (32)$$

which is exactly zero at $\Delta = 0$. Hence by choosing a positive but sufficiently small value for Δ, and pursuing the alternative strategy, the firm can increase its value above the optimal value. Hence $\dot{q}_i^*(t)$ and $r_i^*(t)$ are not (jointly) optimal, a contradiction. This contradiction rules out the possibility that $\pi_i(t) < 0$ and $\dot{q}_i^*(t) \le 0$ at any instant in time.

Conclusion. Combining the results of steps 1-5, $\pi_i(t) \ge 0$ or $\dot{q}_i^*(t) > 0$ (or both) for all t from entry to exit. Also part B of the Lemma was established during step 1. ■

Lemma 7. Suppose that $\dot{p}(t) < 0$ and $p(t)$ is continuous. There is no time interval (a,b), $b > a$, throughout which $q_i^*(t) = 0$.

Proof: Consider first the end of the firm's productive history. If $q_i^*(t) = 0$ were true throughout an interval $(\tau, T_i^*]$, for any $\tau < T_i^*$, then profit would be zero throughout the interval (note R&E spending would be zero using (8)), and hence it would be optimal to exit at any time during the interval. Hence the conditions $\pi_i(t) = 0$ and $\dot{q}_i^*(t) = 0$, required at the exit time from optimal control theory, would have to hold throughout the interval. This is impossible because from Lemma 6B, the conditions for exit, $\pi_i(t) = 0$ and $\dot{q}_i^*(t) = 0$, cannot both occur throughout the interval $(\tau, T_i^*]$ for any $\tau < T_i^*$.

Consider next all other times in the firm's history. During any closed subinterval of (a,b) of length D throughout which $q_i^*(t) = 0$, profit is always zero. Let τ denote the time when the subinterval begins. If the firm moved forward its policies for growth and R&E by D, choosing from the beginning of the period onward $\dot{q}_i(t) = \dot{q}_i^*(t+D)$ and $r_i(t-\delta) = r_i^*(t+D-\delta)$, and exiting at $T_i^* - D$, by Lemma 4B it would enhance its value. This

contradicts the optimality of $\dot{q}_i^*(t)$, indicating that a positive-length time interval with $q_i^*(t) = 0$ throughout does not occur even before the end of the firm's history. ■

Theorem 2. Suppose that $\dot{p}(t) < 0$ and $p(t)$ is continuous. For all t from E_i to T_i^* (excluding T_i^*), $\lambda_i(t) > 0$ and $\dot{q}_i^*(t) > 0$. Also, T_i^* is unique.

Proof: Continual firm growth follows from the preceding Lemmas, as will be seen by analyzing the price-cost margin, $\lambda_i(t)$, $\dot{\lambda}_i(t)$, $\pi_i(t)$, and $\dot{q}_i^*(t)$. Lemma 6A shows that at all t in the firm's production history, either $\pi_i(t) \geq 0$ or $\dot{q}_i^*(t) > 0$. If $\dot{q}_i^*(t) > 0$ then Lemma 2 shows that $\lambda_i(t) > 0$. If $\pi_i(t) \geq 0$ and $q_i^*(t) \neq 0$ then from (3) it must be the case that $p(t) - \overline{c}(t)\left[1 - f\left(\alpha s_i r_i^*(t-\delta)\right)\right] > 0$, and from (13) in Theorem 1D this implies $\dot{\lambda}_i(t) < 0$. Therefore either $\lambda_i(t) > 0$ or $\dot{\lambda}_i(t) < 0$ or $q_i^*(t) = 0$ throughout the firm's productive history. If $q_i^*(t) = 0$, this can last only for an instant in time, not an interval, by Lemma 7, and hence the value of $\lambda_i(t)$ is continuous at times when $q_i^*(t) = 0$ (note $\dot{\lambda}_i(t)$ is finite using (13)). Work backward from time T_i^* when, by Theorem 1D, $\lambda_i(T_i^*) = 0$. The fact that $\lambda_i(t) > 0$ or $\dot{\lambda}_i(t) < 0$ (with $\lambda_i(t)$ continuous at any points in time when $q_i^*(t) = 0$) implies $\lambda_i(t) > 0$ for all previous times. By Lemma 2, this implies $\dot{q}_i^*(t) > 0$ for all $t < T_i^*$.

Since the solution to the Euler equation (9) is unique by Theorem 1B, using Lemma 2 the solution for $\lambda_i(t)$ is unique. Further, Lemma 6B shows that the conditions for exit cannot hold over an interval of positive length, so only at a single t is $\lambda_i(t) = 0$, as required for exit by Theorem 1D. Hence the exit time T_i^* is unique. ■

C. Comparative Firm Behavior

This subsection proves comparative properties of firm behavior. Theorem 3 shows that innovative outputs increase in firm size and skill, and that innovative inputs (R&E spending) increase in firm size but not necessarily firm skill. Theorem 4 shows that among firms of equal

skill, earlier entrants grow faster and produce longer than later entrants. Theorem 5 proves that among simultaneous entrants, more-skilled firms grow faster and produce longer than less-skilled firms.

Theorem 3. For each firm i producing at time t, its R&E output $f(\cdot)$ is higher, and its average production cost is lower, the higher are $q_i(t)$ and s_i. Its R&E spending is higher the higher is $q_i(t)$, but is not necessarily increasing in s_i.

Proof: Substituting the optimal R&E decision from (8) into $f(\cdot)$ yields

$$f\left(\alpha s_i r_i^*(t-\delta)\right) = f\left(\tilde{f}\left(Z_{it}\right)\right), \tag{33}$$

where $Z_{it} = \dfrac{e^{\psi\delta}}{\alpha s_i \bar{c}(t) q_i(t)}$ and $\tilde{f} = f'^{-1}$ are used for compact notation. Note from (8) that $f'\left(\tilde{f}\left(Z_{it}\right)\right) = Z_{it}$ and $\tilde{f}'\left(Z_{it}\right) = 1 \Big/ f''\left(\alpha s_i r_i^*(t-\delta)\right)$.[18] Differentiating (33) with respect to $q_i(t)$ and s_i yields:

$$\frac{\partial f}{\partial q_i(t)} = \frac{-Z_{it}^2}{f''\left(\alpha s_i r_i^*(t-\delta)\right) q_i(t)} > 0, \text{ and} \tag{34}$$

$$\frac{\partial f}{\partial s_i} = \frac{-Z_{it}^2}{f''\left(\alpha s_i r_i^*(t-\delta)\right) s_i} > 0, \tag{35}$$

[18] To see that $f'\left(\tilde{f}\left(Z_{it}\right)\right) = Z_{it}$, start with $f'\left(\alpha s_i r_i^*(t-\delta)\right) = Z_{it}$ as in (8), invert to obtain $\alpha s_i r_i^*(t-\delta) = \tilde{f}(Z_{it})$, compute $f'(\cdot)$ of both sides to obtain $f'\left(\alpha s_i r_i^*(t-\delta)\right) = f'\left(\tilde{f}(Z_{it})\right)$, and note from (8) that the left-hand side is Z_{it}. To see that $\tilde{f}'\left(Z_{it}\right) = 1 \Big/ f''\left(\alpha s_i r_i^*(t-\delta)\right)$, start with $f'\left(\alpha s_i r_i^*(t-\delta)\right) = Z_{it}$ then differentiate with respect to Z_{it} and rearrange to obtain $\alpha s_i \dfrac{\partial r_i^*(t-\delta)}{\partial Z_{it}} = \dfrac{1}{f''\left(\alpha s_i r_i^*(t-\delta)\right)}$; then similarly differentiate both sides of $\tilde{f}(Z_{it}) = \alpha s_i r_i^*(t-\delta)$ to obtain $\tilde{f}'(Z_{it}) = \alpha s_i \dfrac{\partial r_i^*(t-\delta)}{\partial Z_{it}} = \dfrac{1}{f''\left(\alpha s_i r_i^*(t-\delta)\right)}$.

which are both strictly positive since $f''(\cdot) < 0$. Since the average production cost is decreasing in $f(\cdot)$, higher values of $f(\cdot)$ imply lower average cost, and hence (34) and (35) imply that average cost decreases with $q_i(t)$ and s_i. This completes the half of the theorem concerning R&E outputs.

To see how R&E inputs vary with $q_i(t)$ and s_i, solve (8) for $r_i^*(t-\delta)$ and differentiate to obtain

$$\frac{\partial r_i^*(t-\delta)}{\partial q_i(t)} = \frac{-Z_{it}}{f''\left(\alpha s_i r_i^*(t-\delta)\right)\alpha s_i q_i(t)} > 0, \text{ and} \tag{36}$$

$$\frac{\partial r_i^*(t-\delta)}{\partial s_i} = -\frac{r_i^*(t-\delta)}{s_i} + \frac{-Z_{it}}{f''\left(\alpha s_i r_i^*(t-\delta)\right)\alpha s_i^2}. \tag{37}$$

While (36) unambiguously indicates that R&E spending increases in $q_i(t)$, (37) indicates that R&E spending increases (decreases) in s_i if $-f''\left(\alpha s_i r_i^*(t-\delta)\right) \underset{(>)}{<} \frac{e^{\psi\delta}}{\alpha^2 s_i^2 \bar{c}(t) q_i(t) r_i^*(t-\delta)}$. ■

Lemma 8. Suppose that $\dot{p}(t) < 0$ and $p(t)$ is continuous. For firm i pursuing its optimal policies $q_i^*(t)$, $r_i^*(t)$, and T_i^*, let $W_i^* = \int_\tau^{T_i^*} e^{-\psi t}\pi_i(t)\,dt$ denote its accumulated value from τ ($< T_i^*$) to T_i^*. If at τ its output were exogenously increased by $\Delta > 0$, and it chose new decisions optimally thereafter, its value from τ until the time of its exit would strictly exceed W_i^*.

Proof: Let $W_i^*(\Delta) = \int_\tau^{T_i^*(\Delta)} e^{-\psi t}\pi_i(t)\,dt$ given $q_i(\tau) = q_i^*(\tau) + \Delta$ with optimal decisions thereafter (note $W_i^*(0) = W_i^*$). Theorem 2 shows that $\lambda_i(\tau) > 0$. Since $\lambda_i(\tau)$ is the marginal valuation of $q_i(\tau)$, by definition $\left.\frac{\partial W_i^*(\Delta)}{\partial \Delta}\right|_{\Delta=0} = \lambda_i(\tau) > 0$; i.e., increasing Δ at the margin from $\Delta = 0$ increases the firm's remaining value.[19] A hypothetical firm k with $s_k = s_i$ and with output

[19] See for example Kamien and Schwartz (1991, p. 138).

$q_k^*(\tau) = q_i^*(\tau) + \Delta$ would likewise have $\lambda_k(\tau) > 0$. By recursion, $\frac{\partial W_i^*(\Delta)}{\partial \Delta} > 0$ for any $\Delta > 0$, and hence $W_i^*(\Delta) > W_i^*$ for any $\Delta > 0$. ■

Lemma 9. Suppose that $\dot{p}(t) < 0$ and $p(t)$ is continuous. Among firms i and j with $s_i = s_j$, if j exits at time T_j^* but i continues producing, then $q_i^*(T_j^*) > q_j^*(T_j^*)$.

Proof: Since i continues producing, $\int_{T_j^*}^{T_i^*} e^{-\psi t}\pi_i(t)dt > 0$. If $q_i^*(T_j^*) = q_j^*(T_j^*)$ then the firms must be identical and hence j could also earn value $\int_{T_j^*}^{T_i^*} e^{-\psi t}\pi_i(t)dt > 0$ by remaining in production at T_j^* and hence would not exit at T_j^*, so $q_i^*(T_j^*) \neq q_j^*(T_j^*)$. Suppose $q_i^*(T_j^*) < q_j^*(T_j^*)$; then by Lemma 8 firm j could earn strictly greater value than i by remaining in production, and hence j would not exit, a contradiction. This leaves as the only possibility $q_i^*(T_j^*) > q_j^*(T_j^*)$. ■

Theorem 4. Suppose that $\dot{p}(t) < 0$ and $p(t)$ is continuous. Consider firms i and j such that $E_i < E_j$ and $s_i = s_j$. For any time t when both i and j are producing, $\dot{q}_i^*(t) > \dot{q}_j^*(t)$, $q_i^*(t) > q_j^*(t)$, $r_i^*(t-\delta) > r_j^*(t-\delta)$, and $f\left(\alpha s_i r_i^*(t-\delta)\right) > f\left(\alpha s_j r_j^*(t-\delta)\right)$. Also, $T_i^* > T_j^*$.

Proof: These properties are proved by analyzing the time paths $q_i^*(t)$ and $q_j^*(t)$. The path $q_i^*(t)$, and likewise $q_j^*(t)$, is, using Theorem 1B, unique. This implies that the time paths $q_i^*(t)$ and $q_j^*(t)$ never intersect: if $q_i^*(t)$ and $q_j^*(t)$ were to intersect at any time τ, firms i and j would be identical at τ and must follow the same unique time path $q_{ij}^*(t)$ thereafter, and similarly solving backward in time from τ the two firms must have $q_{ij}^*(E_{ij}) = 0$ at the same time E_{ij}, which is both firms' time of entry since $q_i(E_i) = q_j(E_j) = 0$ and since both firms expand continually (Theorem 2), so $E_i = E_j$; this is a contradiction and hence the curves never intersect. Given the lack of intersection and continual expansion, it follows that $q_i^*(t) > q_j^*(t)$. The first

firm to exit must be the smaller firm by Lemma 9, implying $T_i^* > T_j^*$. Since $q_i^*(t) > q_j^*(t)$, by Theorem 3, $r_i^*(t-\delta) > r_j^*(t-\delta)$ and $f\left(\alpha s_i r_i^*(t-\delta)\right) > f\left(\alpha s_j r_j^*(t-\delta)\right)$.

Using Theorem 2 and Theorem 1D, $\lambda_i(T_j^*) > \lambda_j(T_j^*) = 0$. But at times E_i through T_j^*, firms i and j are identical except that $q_i^*(t) > q_j^*(t)$. Differentiating (13) with respect to $q_i(t)$, and noting $\eta_i(t) = 0$ since $q_i^*(t) > 0$ using Theorem 2, yields

$$\frac{d\dot{\lambda}_i(t)}{dq_i(t)} = -e^{-\psi t + \chi(E_i - \delta)} \alpha s_i \bar{c}(t) f'\left(\alpha s_i r_i^*(t-\delta)\right) \frac{\partial r_i^*(t-\delta)}{\partial q_i(t)} < 0, \tag{38}$$

implying that the larger firm always has the greater rate of decrease (or lesser rate of increase), $\dot{\lambda}_i(t) < \dot{\lambda}_j(t)$. Working backward in time from T_j^*, it follows that $\lambda_i(t) > \lambda_j(t)$ for any t when both firms are producing, and from Lemma 2, $\dot{q}_i^*(t) > \dot{q}_j^*(t)$. ■

Theorem 5. Suppose that $\dot{p}(t) < 0$ and $p(t)$ is continuous. Consider firms i and j such that $E_i = E_j$ and $s_i > s_j$. At any time t when both i and j are producing, $\dot{q}_i^*(t) > \dot{q}_j^*(t)$, $q_i^*(t) > q_j^*(t)$, and $f\left(\alpha s_i r_i^*(t-\delta)\right) > f\left(\alpha s_j r_j^*(t-\delta)\right)$. Also, $T_i^* > T_j^*$.

Proof: First it is shown that $q_i^*(t)$ never falls below $q_j^*(t)$ while both i and j are producing. Suppose that $q_i^*(t)$ falls below $q_j^*(t)$ at some time t_1. Let t_2 denote the next time when $q_i^*(t)$ rises to intersect $q_j^*(t)$, or $\min(T_i^*, T_j^*)$ if $q_i^*(t)$ never subsequently intersects $q_j^*(t)$. Since $q_j^*(t)$ is firm j's optimal output path, j earns greater value by choosing curve $q_j^*(t)$ in preference to curve $q_i^*(t)$ throughout the time interval $(t_1, t_2]$. But i could reap an even greater benefit than j during this time interval by choosing $q_j^*(t)$ in preference to $q_i^*(t)$ (and hence $\dot{q}_j^*(t)$ instead of $\dot{q}_i^*(t)$) throughout the period and by choosing optimal R&E spending using (8), because defining $\pi_i^{r^*}(t)$ to be i's effective profit after substituting (8),[20]

[20] To obtain (39), simplify using $f'\left(\tilde{f}\left(Z_{it}\right)\right) = Z_{it}$ and $\tilde{f}'\left(Z_{it}\right) = 1 \Big/ f''\left(\alpha s_i r_i^*(t-\delta)\right)$ as in Theorem 3.

$$\frac{d}{ds_i}\left(\frac{d\pi_i^{r^*}(t)}{dq_i(t)}\right)=\frac{-e^{\chi(E_i-\delta)+2\psi\delta}}{\alpha^2 s_i^3 \bar{c}(t) f''\left(\alpha s_i r_i^*(t-\delta)\right) q^2(t)}>0 \tag{39}$$

and because i's and j's growth costs are identical if they choose the same curve. If $q_i^*(t)$ rises to intersect $q_j^*(t)$, then i could therefore increase its value by choosing output $q_j^*(t)$ instead of $q_i^*(t)$ and choosing optimal R&E spending accordingly throughout the interval $(t_1,t_2]$. If $q_i^*(t)$ does not rise to intersect $q_j^*(t)$ and $T_i^* \le T_j^*$, then i could therefore increase its integrated discounted profit over the interval $(t_1,t_2]$ by choosing output $q_j^*(t)$ instead of $q_i^*(t)$ and choosing optimal R&E spending accordingly throughout that interval. If $q_i^*(t)$ does not rise to intersect $q_j^*(t)$ and $T_i^* > T_j^*$, then i could therefore increase its integrated discounted profit over the interval $(t_1,t_2]$ by choosing output $q_j^*(t)$ instead of $q_i^*(t)$ and choosing optimal R&E spending accordingly throughout that interval, and Lemma 8 shows that as a result of having output $q_j^*(t_2) > q_i^*(t_2)$ at time t_2 i would also gain greater integrated discounted profit thereafter. Hence in each of these three cases, $q_i^*(t)$ must not be optimal, a contradiction, proving that $q_i^*(t)$ never falls below $q_j^*(t)$.

Next it is shown that $T_i^* > T_j^*$, via implications of exit time for value given skill and output of firms i and j and hypothetical firm k. Suppose that $T_i^* \le T_j^*$. If j exits at $T_j^* > T_i^*$, j must earn positive discounted profits over the time interval $(T_i^*,T_j^*]$; let $W_j^* = \int_{T_i^*}^{T_j^*} e^{-\psi t}\pi_j(t)\,dt$ denote these discounted profits. $W_j^* > 0$ if $T_i^* < T_j^*$, or $W_j^* = 0$ if $T_i^* = T_j^*$. Consider a firm k with skill $s_k = s_j$ and output $q_k(T_i^*) = q_i^*(T_i^*) \ge q_j^*(T_i^*)$ at time T_i^*, and let W_k^* denote the maximum discounted profit k could accumulate from T_i^* onward and P_k^* denote the set of policies (for growth and R&E spending from T_i^* until exit, and for exit time) that yield this maximum profit. $W_k^* > W_j^*$ if $q_i^*(T_i^*) > q_j^*(T_i^*)$, by Lemma 8, or $W_k^* = W_j^*$ if $q_i^*(T_i^*) = q_j^*(T_i^*)$. Using equation (3), since $s_i > s_j$, firm i could earn even greater discounted profits than W_k^* by, instead of exiting at T_i^*, choosing the set of policies P_k^*; letting $\tilde{W}_i$ denote i's resulting

discounted profits, $\tilde{W}_i > W_k^*$. Since $\tilde{W}_i > W_k^* \geq W_j^* \geq 0$, i could increase its value by exiting later than T_i^* and choosing appropriate policies thereafter. This indicates that T_i^* is not optimal, a contradiction. Hence it is impossible that $T_i^* \leq T_j^*$.

This leaves only the possibility that $q_i^*(t) \geq q_j^*(t)$ and $T_i^* > T_j^*$, and further insights result by examining $\lambda_i(t)$ and $\lambda_j(t)$. Equation (13) dictates the changes of $\lambda_i(t)$ and (replacing i with j) of $\lambda_j(t)$. Differentiating (13) with respect to size and skill, which are the only differences between i and j at each $t \leq T_j^*$, yields $\frac{\partial \dot{\lambda}_i(t)}{\partial q_i(t)} < 0$ as indicated by (38) and $\frac{\partial \dot{\lambda}_i(t)}{\partial s_i} < 0$.[21] Since i is more skilled and at least as large as j, it follows that $\dot{\lambda}_i(t) < \dot{\lambda}_j(t)$ for all t from the time of their entry to firm j's exit. Moreover, $\lambda_i(T_j^*) > \lambda_j(T_j^*) = 0$ using Theorem 2 and Theorem 1D. Working backward in time from when each firm exits, it follows that $\lambda_i(t) > \lambda_j(t)$ for any t when both firms are producing, and from Lemma 2, $\dot{q}_i^*(t) > \dot{q}_j^*(t)$, and hence $q_i^*(t) > q_j^*(t)$. Since $q_i^*(t) > q_j^*(t)$ and $s_i > s_j$, Theorem 3 indicates that $f\left(\alpha s_i r_i^*(t-\delta)\right) > f\left(\alpha s_j r_j^*(t-\delta)\right)$. ■

D. Industry-Wide Outcomes

Theorems in this subsection demonstrate outcomes that pertain to the whole industry. Theorem 8 shows that the assumed equilibrium exists in which industry output steadily increases and the price charged to consumers steadily falls, with both output and price continuous. Theorem 9 shows that entry eventually ceases and, after an initial increase in the number of firms and decrease in concentration, the number of firms eventually declines steadily while concentration increases.

Profit opportunities are high for potential entrants in the early years of the industry, but a falling price increasingly lowers future profit opportunities. Eventually even the most skilled

[21] To prove that $\frac{\partial \dot{\lambda}_i(t)}{\partial s_i} < 0$, differentiate (13) with respect to s_i, noting from Theorem 3 that $\frac{\partial f\left(\alpha s_i r_i^*(t-\delta)\right)}{\partial s_i} > 0$.

potential entrants can no longer earn positive value by entering, and entry ceases. Incumbents continually expand, with the largest and most-skilled firms expanding fastest. Increasingly the market becomes dominated by a few early-entering high skilled firms, while other producers are steadily driven out of business. Consecutive cohorts of entrants are driven to extinction in reverse order of entry.

Figure 3 illustrates the situation after entry has ceased, at a time τ. No firms have size 0, because all surviving firms have grown. Surviving firms have the sizes and skills indicated by the shaded region. Firms with the lowest size and skill have the least profit and are first to exit. These firms are indicated by the dashed curve labeled "margin of survival," and at τ they earn zero profit and have no growth. Firms below the margin of survival may exist until a short while after entry ceases, as they race to grow large enough to attain the shaded region and earn a positive profit. The top of the shaded region is bounded by the maximum possible skill, $\overline{s}$. The right hand curve of the shaded region consists of firms that entered at the earliest possible time, 0. The largest and most-skilled firms have the highest growth rate, and hence the highest growth rate of all is attained by firms at the top-right of the shaded region. Within any entry cohort such as the entrants at E_i, since the most-skilled firms have always had the highest growth rate, they are largest by time τ. Over time following τ, the margin of survival shifts to the upper right, and the larger and more-skilled firms continue to have higher growth than other firms, increasing industry market share among increasingly few firms.

Lemma 10: The margin of survival, along which $\pi_i(t) = 0$ for firms with $\dot{q}_i(t) = 0$, is

$$s_i = k(t)/q_i(t), \tag{40}$$

where $k(t)$ satisfies

$$\alpha k(t)\left(p(t) - \overline{c}(t)\left[1 - f\left(\tilde{f}\left(\frac{e^{\psi\delta}}{\alpha k(t)\overline{c}(t)}\right)\right)\right]\right) = e^{\psi\delta}\,\tilde{f}\left(\frac{e^{\psi\delta}}{\alpha k(t)\overline{c}(t)}\right), \tag{41}$$

for all firms i exiting at time t. (This holds for any $p(t)$ regardless whether $p(t)$ is strictly decreasing.)

Proof: Start with $\pi_i(t) = 0$ and substitute using the optimal R&E equation (8) and the terminal conditions (10) and (11). Use the implicit function theorem to find $ds_i / dq_i(t)$, then solve the resulting differential equation for s_i as a function of $q_i(t)$, yielding (40). Substitute (40) into $\pi_i(t) = 0$ to reveal how $k(t)$ is determined at each t: it is the solution to (41). ■

Lemma 11: Write output growth at time t (defined in the same manner as a derivative but with no guarantee that it need be finite) as

$$\dot{Q}(t) = \dot{Q}_F + \dot{Q}_X + \dot{Q}_R + \dot{Q}_C, \tag{42}$$

where $\dot{Q}_F$ is output growth by incumbent firms, $\dot{Q}_X \leq 0$ is output growth from intentional exit, $\dot{Q}_R < 0$ is output growth from random exit, and $\dot{Q}_C \geq 0$ is output growth from international competitors. Let $w(t;s,q)$ denote the density (number per unit of skill and output) at time t of firms with skill s and output q. If $w(t;s,q)$ is finite for all s and q, then $\dot{Q}(t)$ can be written

$$\dot{Q}(t) = \iint_{\Omega} \dot{q}^*(t;s,q)\, w(t;s,q)\, ds\, dq - \max\left(0, \int_{s_a}^{s_b} \frac{k(t)}{s} w\left(t;s,\frac{k(t)}{s}\right)\frac{\dot{k}(t)}{s}\, ds\right)$$
$$- \chi \iint_{\Omega} q^*(t;s,q)\, w(t;s,q)\, ds\, dq + \dot{Q}_C(t), \tag{43}$$

where Ω denotes the set of combinations of skill s and output q for which there are non-exiting firms, $q^*(t;s,q)$ denotes the output at time t of each firm with skill s and output q, and s_a and s_b respectively denote the lowest and highest skill of surviving firms that are about to exit (those on the margin of survival that are not racing for profitability; $s_b \leq \bar{s}$). (This holds for any $p(t)$ regardless whether $p(t)$ is strictly decreasing.)

Proof: Expression (42) simply cumulates the sources of growth and exit in the model (recall that entrants begin with zero output so entry at t does not impact $\dot{Q}(t)$). If $w(t;s,q)$ is finite, consider each term in (42). Output growth by incumbents is

$$\iint_{\Omega} \dot{q}^*(t;s,q)\, w(t;s,q)\, ds\, dq, \tag{44}$$

Expression (44) is simply the integral, over the area of Figure 3 that is shaded plus any area in which firms are racing to attain profitability, of the (absolute) growth rates of these firms.

Output loss by intentional exit occurs as the margin of survival in Figure 3 sweeps across the bottom left region of the shaded part of the figure; the firms swept across (unless they are still racing for profitability) choose to exit. From (40), the margin of exit consists of all points (s,q) that satisfy $q = k(t)/s$, and the motion over time of each point in the margin of survival can be described by $\left(\frac{ds}{dt}, \frac{dq}{dt}\right) = \left(0, \frac{\dot{k}(t)}{s}\right)$. The quantity of firms exiting per unit of time is therefore a flux, or surface integral, across the (one-dimensional) surface corresponding to the current margin of survival (from s_a to s_b). The flow across that surface equals the motion of the margin of survival times the quantity of firms per unit of skill and output at each point (s,q) on the surface. The surface integral for output loss simplifies to

$$\int_{s_a}^{s_b} \frac{k(t)}{s} w\left(t; s, \frac{k(t)}{s}\right) \frac{\dot{k}(t)}{s}\, ds\,. \tag{45}$$

In this integral the term $\frac{k(t)}{s}$ is the value of q corresponding to s at each point on the margin of survival, and the term $\frac{\dot{k}(t)}{s}$ is the horizontal motion of the margin of survival. The integral correctly describes output loss due to exit if $\dot{k}(t) \geq 0$. If $\dot{k}(t) < 0$ the margin of survival is moving backward and the integral is a negative number, the absolute value of which is the quantity of firms per unit of time that match the requirement for exit $\pi_i(t) = 0$. A backward shift in the margin of survival corresponds to growth in profit, $\dot{\pi}_i(t) > 0$, even for a firm that chooses $\dot{q}_i(t) = 0$ and hence no firm exits in this situation as it can earn positive profit by remaining in the industry longer. Hence output loss equals (45) if $\dot{k}(t) \geq 0$ or 0 if $\dot{k}(t) < 0$, i.e.,

$$\max\left(0, \int_{s_a}^{s_b} \frac{k(t)}{s} w\left(t; s, \frac{k(t)}{s}\right) \frac{\dot{k}(t)}{s}\, ds\right) \tag{46}$$

for intentionally exiting firms.

Output loss by random exit equals the random exit rate χ times the total quantity being produced by firms not intentionally exiting, i.e.,

$$\chi \iint_{\Omega} q^*(t; s, q)\, w(t; s, q)\, ds\, dq\,. \tag{47}$$

The net output gain per unit of time at t, $\dot{Q}(t)$, is therefore output growth (44), less output loss due to intentional exit (46), less output loss due to random exit (47), plus output growth by international competitors $\dot{Q}_C(t)$, yielding (43). ■

Lemma 12: $p(t)$ is continuous and differentiable. Also, $q_i^*(t)$ and $\dot{q}_i^*(t)$ are always finite for all i. (This holds for any $p(t)$ regardless whether $p(t)$ is strictly decreasing.)

Proof: By assumption $D^{-1}\left(Q(t);t\right)$ is C^1, so $p(t) = D^{-1}\left(Q(t);t\right)$ is continuously differentiable if $Q(t)$ is continuous with a finite derivative. This is assured if (42) is finite, so consider the four terms in (42) in the following order: growth (first term), random exit (third term), international competition (fourth term), and intentional exit (second term). The proof does not make use of Lemma 11, only the change in $Q(t)$ given by equation (42).

The growth term is finite because (a) there is a finite number of firms and (b) every firm has finite growth. (a) An upper bound to the number of firms at t is $\int_0^t P(\tau)\,d\tau$, which is finite since potential entry $P(t)$ is finite. (b) Infinite growth at a time τ, $\dot{q}_i(\tau) = \infty$ or $\dot{q}_i(\tau) = -\infty$, corresponds to a step change $\Delta q \neq 0$ in $q_i(t)$. It will be shown that this cannot be optimal, by considering the minimum cost of growth by Δq in a time interval, then comparing this cost to an upper bound on market-wide revenues. The discounted cost of growing by Δq in a brief interval of length Δt is at least $e^{-\psi(\tau+\Delta t)} g\left(\frac{\Delta q}{\Delta t}\right)\Delta t$. Starting with any $\Delta t > 0$, say $\Delta t = 1$ year, consider how halving Δt would affect $g\left(\frac{\Delta q}{\Delta t}\right)\Delta t$. At the starting value of Δt, $g\left(\frac{\Delta q}{\Delta t}\right)\Delta t > 0$. At the first halving, $g\left(\frac{\Delta q}{\Delta t}\right)$ increases by a multiple $\omega_1 > 2$. At the ℓ^{th} halving, for $\ell \geq 2$, $g\left(\frac{\Delta q}{\Delta t}\right)$ increases by a multiple $\omega_\ell > 2\omega_{\ell-1}$. (This can easily be seen by graphing $g(\cdot)$ and noting that the argument $\frac{\Delta q}{\Delta t}$ doubles each time.) Hence at Δt's first halving $g\left(\frac{\Delta q}{\Delta t}\right)\Delta t$ increases by a multiple $\frac{\omega_1}{2} > 1$, and at its ℓ^{th} halving ($\ell \geq 2$) by a multiple $\frac{\omega_\ell}{2} > \frac{2\omega_{\ell-1}}{2} = \omega_{\ell-1}$. After

undergoing the ℓ^{th} halving of Δt, $g\left(\frac{\Delta q}{\Delta t}\right)$ has grown by a multiple $\prod_{m=1}^{\ell}\omega_m > \sigma^{\ell}$ where $\sigma = \omega_1/2 > 1$. Choose any ℓ such that $e^{-\psi(\tau+1\text{ year})}g\left(\frac{\Delta q}{1\text{ year}}\right)(1\text{ year})\sigma^{\ell} > \overline{V}$, where $\overline{V}$ is an upper bound on the market value obtained by all firms combined, $\overline{V} = \int_0^{\infty} e^{-\psi t}\,\overline{R}\,dt = \overline{R}\int_0^{\infty} e^{-\psi t}dt$ with $\overline{R}$ being the (assumed) upper bound on total market revenue $D^{-1}(Q;t)\cdot Q$. Since τ lies within a time interval of length $\Delta t = 2^{-\ell}\,year$, the cost of growing by Δq at τ must exceed $\overline{V}$. Hence no firm, not even all firms combined, could earn enough revenue to pay the cost of so much growth in such a short interval. Firm i would earn greater value by never entering, so choosing $\dot{q}_i(\tau) = \infty$ or $\dot{q}_i(\tau) = -\infty$ cannot be optimal. This concludes the proof that the growth term is finite.

The random exit term is finite because (a) there is a finite number of firms, (b) every firm has finite output, and (c) they exit randomly at finite rate χ. Point (a) was shown when analyzing the growth term, (b) is a consequence of finite firm growth as shown when analyzing the growth term, and (c) was assumed.

The international competition term in (42), $\dot{Q}_C(t)$, is finite by assumption.

The intentional exit term is also finite. If exit were infinite, while the other terms in (42) are finite as has been shown, the absolute increase in price would be strictly positive and every firm that exited could have earned positive profit by remaining in production, and hence would not have exited, a contradiction.

Since each of the four terms in (42) is finite, (42) is finite.

Above it was shown that $\dot{q}_i^*(t)$ is finite, which implies that $q_i^*(t) = \int_{E_i}^{t}\dot{q}_i^*(\tau)\,d\tau$ is finite.

■

Lemma 13: If $\dot{p}(t) \geq 0$ at any t, with $p(t)$ continuous and differentiable, and if $\dot{q}_i^*(t) \geq 0$ for all i, with $w(t;s,q)$ finite, then no firms intentionally exit at t.

Proof: $\dot{p}(t) \geq 0$ implies that the margin of survival stays the same or shifts downward, $k'(t) \leq 0$. Using this fact, $\dot{Q}_X = -\max\left(0, \int_{s_a}^{s_b} \frac{k(t)}{s} w\left(t; s, \frac{k(t)}{s}\right) \frac{\dot{k}(t)}{s}\, ds\right)$ from Lemma 11 is exactly zero at t. ■

Theorem 8: An equilibrium exists satisfying the assumption that $p(t)$ is continuous with $\dot{p} < 0$ for all t, for a sufficiently small value of χ (including positive values). Also, $Q(t)$ is continuous with $\dot{Q} > 0$ for all t.

Proof: It has been assumed that $p(t)$ is strictly decreasing, and Lemma 12 shows that $p(t)$ is continuous and differentiable, so $\dot{p} < 0$. By Theorem 2, hence, $\dot{q}_i^*(t) > 0$ for all firms with $E_i \leq t < T_i^*$. Moreover, with earlier-entering and more-skilled firms growing faster as established in Theorems 4 and 5, and with $P(t)$ and $H(s_i)$ finite, the density $w(t; s, q)$ of firms with respect to size and skill remains finite for each combination of entry time and skill (E_i, s_i), and the fraction of firms on the margin of survival is of zero measure.

We wish to prove that for some choice of χ, firm behavior *given* $\dot{p} < 0$ can yield aggregate expansion of output, $\dot{Q} > 0$, and hence is consistent with the assumption $\dot{p} < 0$. We will show that the firm behavior in fact must yield aggregate expansion of output, by supposing the opposite, working back from time κ, and examining $\dot{Q}$ and its terms in (43). Suppose $\dot{Q}(t) \leq 0$ at some time t. Recall that $\dot{Q} > 0$ for all $t > \kappa$, and work backward in time to the last previous time τ when $\dot{Q} \leq 0$. By continuity of p, $\dot{Q}(\tau) = 0$. The first (firm expansion) term in (43) is strictly positive because, as established above, $\dot{q}^*(t; s, q) > 0$ for all non-exiting firms. The second (intentional exit) term is zero, since $\dot{Q}(t) \leq 0$ implies $\dot{p}(t) \geq 0$, using Lemma 13. Hence choosing

$$\chi < \frac{\iint_{\Omega} \dot{q}^*(\tau; s, q)\, w(\tau; s, q)\, ds\, dq + \dot{Q}_C(\tau)}{\iint_{\Omega} q^*(\tau; s, q)\, w(\tau; s, q)\, ds\, dq} \tag{48}$$

ensures $\dot{Q}(\tau) > 0$, which violates the supposition and hence proves that firm behavior given $\dot{p} < 0$ must satisfy $\dot{Q} > 0$, and hence that the equilibrium exists.

As noted above, using Lemma 12, $p(t)$ is continuous and differentiable and $\dot{p} < 0$, so $Q(t)$ is continuous with $\dot{Q} > 0$ using the inverse demand curve. ■

Theorem 9. Entry eventually ceases. The number of firms increases initially, but eventually strictly declines over time until only the highest-skilled earliest entrants remain in production. Concentration of market shares, although initially high and strictly decreasing, eventually is generally increasing as earlier-entering and more-skilled firms steadily take over the market.

Proof: Let i be a relatively early (but not earliest) entrant with skill s_i; it exits at time T_i^*. Compare i versus a potential entrant j with the same skill, $s_j = s_i$, that chooses whether to enter at time $E_j \geq T_i^*$. If j enters it must go through a time of negative profit, because $\pi_j(t) < \pi_i(T_i^*)$ for $t \geq E_j$ and $q_j^*(t) < q_i^*(T_i^*)$, before the time τ when it attains the size $q_i^*(T_i^*)$ that i had when it exited. Moreover, once it attains that size, the value it accumulates from then onward must be negative, $\int_\tau^{T_j^*} e^{-\psi t}\pi_j(t)dt < 0$: if j could achieve non-negative value from τ onward, then by Lemma 4 firm i would have been able to obtain strictly positive value by continuing to produce at time T_i^* and hence i would not have exited when it did. Hence the total value that j could accumulate by entering at E_i and producing must be strictly negative, so j never enters production. This proof pertains to firms at any level of skill up to the maximum, $\bar{s}$, so eventually even potential entrants of maximum skill stop entering.

The number of firms initially is increasing from zero. At time t the output of all firms is $\int_0^\infty \int_{-\infty}^{\bar{s}} h^d(E,s;t)q^d(E,s;t)ds\,dE$, where E represents entry time; s represents skill; $h^d(E,s;t)$ represents the number of producing firms per unit of E and s, as a function of E, s, and t; and $q^d(E,s;t)$ represents the output of firms as a function of E, s, and t. The output of only the first entrants, $E = 0$, is $\int_0^0 \int_{-\infty}^{\bar{s}} h^d(E,s;t)q^d(E,s;t)ds\,dE = 0$; i.e., their output is negligible (of measure zero) relative to the output of all firms, so concentration must decrease initially.

Once entry ceases, the number of firms can only decrease as firms exit. And in fact the number of firms then must be strictly decreasing at every t, as some exit is always taking place. To prove this, suppose that no exit took place over a period of time from t_1 to t_2 and let j and i denote firms of equal skill that exit at t_1 and t_2 respectively. At time $t_1 = T_j^*$, j and i differ only infinitesimally in output (for any $\varepsilon > 0$, $q_i^*(t) - q_j^*(t) < \varepsilon$), so by the uniqueness of T_i^* established in Theorem 2 (and the continuity of (12) in Theorem 1D), i must earn positive value from t_1 to t_2. By paying an infinitesimal cost to grow slightly instead of choosing $\dot{q}_j^*(T_j^*) = 0$, firm j could earn more value than i earns, and more than the cost of growth, so j would not have exited at time t_1. Hence with some exit always occurring, the number of firms is strictly decreasing over time. Moreover, since by Theorems 4 and 5 the earliest-entering and most skilled firms always expand most rapidly, these firms gain increasing market share. Since relatively small, high-skill firms may grow faster than larger, lower-skill firms, it is conceivable that industry concentration could decrease at some times, but in general measures of market concentration must tend to increase as output is concentrated among increasingly few firms. ■

E. Inter-Industry Differences

This subsection shows the general nature of inter-industry differences caused by variation in α, the potential for R&E to reduce cost or improve quality. The precise effects of α depend on the functional forms of the model, and theorems in this section have not been proved to pertain to every gradation of α at every t. Nonetheless, several important points can be made that characterize the general effects of α on firm and industry outcomes. Theorem 10 proves that in the limit as α approaches zero, key results of the model break down: firms spend no money on R&E, and firms' size and skill – and hence entry time as well – are irrelevant to growth and exit. Moreover, Theorem 10 points out that if there is demand growth, or if firms have any risk of random exit for reasons such as management failures or natural disasters, then as long as international competitors do not account for the requisite new production, entry never ceases. Similarly with any growth in product demand, for α very low, continued entry would occur.

More generally, competition tends to be more severe, and shakeouts more rapid, given any increase in α. Theorem 11 proves formally that the price and cost curves, $p(t)$ and $\bar{c}(t)$,

are affected by α. The resultant effects are complicated: lower future prices mean less incentive to expand, partially counteracting the greater expansion caused by α being higher. Therefore no unambiguous proof is shown of its effects at all t. However, the general nature of the effect is clear: because firms have more incentive to do R&E and to grow, the competitive advantage of earlier-entering (hence having more time to grow) and more-skilled firms is amplified. With a fiercer competitive environment, entry tends to be foreclosed relatively early, forcing a shakeout in the number of firms to begin by an earlier date. The difference between cohorts in terms of growth, exit time, and R&E spending and output is hence increased.

Theorem 10. In the limit as α approaches zero, R&E has no impact on firms' costs; firms spend zero money on R&E; and size and skill have no effect on firm growth or exit. If $\frac{\partial D^{-1}(Q;t)}{\partial t} > 0$ or $\chi > 0$ (or both), such that $\dot{Q}_D + (-\dot{Q}_R) > \dot{Q}_C$ where Q_D is market demand at the equilibrium price, then entry continues forever.

Proof: The assumptions about $f(\cdot)$ imply that, in the limit as α approaches zero, optimal R&E spending is $\lim_{\alpha\to 0} r_i^*(t-\delta) = \frac{1}{\alpha s_i} f'^{-1}\left(\frac{e^{\psi\delta}}{\alpha s_i \bar{c}(t) q_i(t)}\right) = 0$. The impact on firms' costs is given by $f(\cdot)$, and $\lim_{\alpha\to 0} f\left(\alpha s_i r_i^*(t-\delta)\right) = f(0) = 0$. Effective profit is then $\lim_{\alpha\to 0} \pi_i(t) = e^{\chi(E_i-\delta)}\left(\left[p(t)-\bar{c}(t)\right]q_i(t) - g\left(\dot{q}_i(t)\right)\right)$. The time path for optimal growth is then given by, using (12) and (13) in Theorem 1D, $\dot{q}_i^*(t) = g'^{-1}\left(e^{\psi t-\chi(E_i-\delta)}\lambda_i(t)\right)$ and $\dot{\lambda}_i(t) = -e^{-\psi t+\chi(E_i-\delta)}\left(p(t)-\bar{c}(t)\right) - \eta(t)$ with the exit requirements being $\dot{q}_i^*(t) = 0$ and $p(t)-\bar{c}(t) = 0$. These equations are independent of both s_i and $q_i(t)$, which hence have no effect on growth or exit. Since the profit function is the same for all firms, the number of firms converges to an equilibrium in which profits are zero and there is no entry or exit. Were there no risk of random exit, no demand growth, and no growth of international competition, the number of firms would stabilize with a profit flow equal to zero. With demand growth yielding $\dot{Q}_D = \dot{Q} > 0$, or continued exit at rate χ that gives total production losses of $-\dot{Q}_R > 0$,

equilibrium requires continual new production by in-country or international firms, and $\dot{Q}_D + (-\dot{Q}_R) > \dot{Q}_C$ implies that some or all of the new production comes from in-country firms. Given that (12) and (13) are independent of firm size, new entrants have identical incentives to incumbents to expand, and hence expansion is not limited to incumbents; entry continues forever. ■

Theorem 11. The price and cost curves, $p(t)$ and $\bar{c}(t)$, are affected by α.

Proof: By symmetry of α and s_i in the model, Theorem 5 shows that *if* $p(t)$ and $\bar{c}(t)$ are held constant, all firms spend more on R&E and expand faster when α is increased from α_1 to α_2. The price and cost curves, $p(t)$ and $\bar{c}(t)$, therefore cannot be unaffected by α, because if they were not, at all t every firm would expand faster with α_2 than with α_1, causing $Q(t)$ to be higher and $p(t)$ to be lower, a contradiction. ■

IV. Testable Implications

The model has several strong, testable implications for cross-industry differences in the dynamics of competition. The differences occur as a result of variations in α, the potential for R&E to reduce cost or improve quality. Three means will be used to test the theory. First is to look across industries, to see whether outcomes predicted for high-α industries occur simultaneously, with industries that do not have these characteristics simultaneously sharing the predicted outcomes for low-α industries. Second is to look across countries at the same industries, to help confirm that underlying product characteristics (such as α) are the cause of industry outcomes rather than random processes of competition or national environments.[22] Third is to use measures of technological change to assess whether industries with more severe shakeouts indeed have more rapid within-industry technological progress (without too-rapid technology diffusion) and which firms are responsible for the technological progress.

The implications of the model hold for industries defined at a competitive level. The automotive industry, for example, is too broad a category, as most firms in the industry produce

[22] On the effects of national environments on industry, see for example Lundvall (1992) and Nelson (1993).

components that do not compete with each other. Using too-aggregated data, including most 4-digit SIC industries, mixes the outcomes of different industries and, unless nearly all sub-industries follow identical dynamics, makes it impossible to observe their competitive dynamics.

Comparisons across countries hold for capitalist economies in which an indigenous industry producing the product develops near the worldwide commercial inception of the product.[23] If different nations' industries were entirely isolated, a test could in theory be carried out to distinguish the role of product market specific causes of dynamics versus random market outcomes and differences in national environments. In practice the industries examined here are not so isolated across countries, and international competition seems typically to have been substantial in the later decades of most products' histories although relatively unimportant during the earlier decades. Nonetheless, especially in the earlier decades, inter-country differences have the opportunity to arise from random causes or national environments, yielding different patterns of entry, initial modes of competition, times with peak numbers of producers, or fractions of weaker firms that might exit during later competition.

Predictions 1 and 2 pertain to the number of firms as it changes over time:

1. The same product industries in different countries experience similar degrees of shakeout.
2. The same product industries in different countries experience similar timing as to when they achieve their peak number of firms.

Since the underlying characteristics that cause industry outcomes are assumed to be predictable and hence the same across nations, outcomes are driven solely by these characteristics and not by random events nor by national environments. In particular, higher R&E potential is associated with an earlier and more rapid shakeout of firms.

Predictions 3 and 4 pertain to the entry and exit processes causing changes in the numbers of firms:

3. In industries with strong shakeouts, entry eventually ceases, or in practice declines considerably since industry data may be contaminated with a trace of firms that serve

[23] Comparisons might also hold for planned economies, if economic realities force the planners to make decisions approximating those of capitalist firms and if in practice much R&E must occur within individual productive units and diffusion across units is difficult.

specialist markets or have motivations outside the theory. With little or no shakeout, entry continues relatively unabated. Similar entry patterns occur in all countries.

4. In industries with strong shakeouts, firms in earlier-entering cohorts tend to exit less frequently than firms in later-entering cohorts, and entry cohorts go extinct in reverse order of entry. With little or no shakeout, these differences are insignificant or nonexistent.

Less entry per year or greater percentage exit per year each could cause a drop in the number of firms. In the theory it is the decrease in entry that guarantees a shakeout, and the annual percentage of firms exiting the industry could remain steady. It is expected that, as suggested in Theorem 10, some exit occurs for random reasons outside the theory. If so, then as long as incumbent firms grow quickly enough to make up for the exit, entry still ceases and shakeout occurs in industries with strong R&E potential, but entry continues and no shakeout occurs in industries with low R&E potential.

Predictions 5 through 7 pertain to the potential for R&E and inter-cohort differences in firm R&E activities:

5. In industries with strong shakeouts, improvements in the product and manufacturing methods tend to be more rapid than in industries without shakeouts.[24]
6. In industries with strong shakeouts, relatively early entrants have the greatest R&E output.
7. In industries with strong shakeouts, R&E output is associated with enhanced probability of firm survival.

A greater potential for R&E causes shakeouts in the model because it yields substantial R&E output (prediction 5) that is dominated by the largest producers, which predominantly entered earlier (prediction 6).[25] And since greater R&E output is associated with size-and-skill

[24] The expression "tend to be" is necessary because in some cases rapid improvements can be purchased through third-party suppliers, as for styrene manufacturers buying process machinery from third-party suppliers, or diffusion of technological information may be easy and rapid.

[25] The model also implies that with little or no shakeout, entrants from all eras are similar in their near-zero R&E output. Unfortunately, the sample sizes involved should then make it virtually impossible to test statistically for similarity in amounts of R&E output. Two other complicating

combinations that mark the more profitable firms, indeed is the very source of the greater profitability of these firms, the successful performers of R&E have the greatest probability of survival.

Two other predictions are also worth noting. First, although R&E output rises with firm size and skill, R&E *intensity* (R&E output per unit of firm size) does not necessarily rise with size and skill. Since greater skill results in greater size, this is loosely interpreted as larger firms having greater R&E output but not necessarily greater R&E intensity. Second, the mean value of results from each dollar of R&E falls with firm size and skill; that is, the marginal value of R&E is decreasing, $f'\left(\alpha s_i r_i^*(t-\delta)\right) < 0$, as R&E increases. These predictions explain, in the same manner as Cohen and Klepper (1996), findings of the cross-sectional empirical literature on the so-called Schumpeterian hypothesis regarding firm size and innovation (Cohen, 1995).

V. Extensions

A. Oligopoly

With an industry that tends toward increasingly concentrated market structure, it is natural to ask how oligopolistic rather than price-taking behavior might influence industry outcomes. Large firms' decisions might deviate from those of their price-taking brethren. To investigate this issue simply in an extreme case, suppose that at time t^m there has arisen a monopolist with a fringe of competitors. Denote the total output of the competitors at time t as a functional $Q_{-m}(p(t);t)$, which depends on the response of the fringe firms to the time path of

---

factors arise in testing predictions 5 and 6. First is the potential for product development that does not qualify as R&E, or for R&E that can be successfully sold between firms. If the incentives for product development and saleable R&E are similar in shakeout and non-shakeout industries, R&E data that are polluted with development and saleable R&E will tend to suggest similar output for earlier and later entrants. Second is inter-industry differences in propensity to patent or record an innovation, or in how dramatic an innovation needs to be before it is patented or recorded as an innovation, which add substantial noise to inter-industry comparisons regarding prediction 5.

price. Assume that the monopolist can be characterized exactly analogously to the model of firms given earlier.[26]

Now total industry output is $Q(t) = Q_{-m}(p(t);t) + q_m(t)$, where the subscript $i = m$ is used to indicate the monopolist. The model could be extended to analyze the behavior of the monopolist in this case and compare its behavior to perfectly competitive behavior, thus providing insights into any differences that might be caused by strategic behavior.

B. Product versus Process R&E

Although R&E has been portrayed above solely for efficiency improvement, the industry dynamics portrayed here are though to stem from not only process but also product R&E. Product R&E enhances quality, which makes some customers willing to pay more and hence allows the firm to obtain a higher mean price than would otherwise be possible at the current point on the marketwide demand curve. The effect of product R&E can be incorporated in the model much as for process R&E, with, it is expected, analogous results. Writing product R&E spending as $r_i^d(t-\delta)$, assume it multiplies the mean price received by a factor $f^d(\alpha s_i r_i^d(t-\delta))$. Here $f^d(\cdot)$ has properties identical to $f(\cdot)$ except that the range of $f^d(\cdot)$ may exceed 1. The firm's optimal choice of product R&E spending then satisfies (8) with $\bar{c}(t)$ replaced by $p(t)$, comparative product R&E patterns are anticipated to be identical to those reported in Theorem 3, and outcomes of the model are anticipated to be identical to those already reported.

VI. Conclusions

This paper has developed a model of industry dynamics with distinctive testable implications. Comparison with empirical fact has strongly supported the ability of the model to explain real-world dynamics across the majority of product-level industries.

[26] The monopolist has finite output, so technically many of the variables and functions applicable to the monopolist differ from those of fringe firms; their units of measurement differ. Were they written down, the fringe firms' variables and functions should be denoted distinctly.

References

Chandler, Alfred D., Jr. *Scale and Scope: The Dynamics of Industrial Capitalism*. Cambridge, MA: Harvard University Press, 1990.

Dasgupta, Partha and Joseph Stiglitz. "Industrial Structure and the Nature of Innovative Activity." *Economic Journal* 90 no. 358, 1980, pp. 266-293.

Flaherty, M. Thérèse. "Industry Structure and Cost-Reducing Investment." *Econometrica* 48, 1980, pp. 1187-1209.

Jovanovic, Boyan and Glenn MacDonald. "The Life Cycle of a Competitive Industry." *Journal of Political Economy*, *102* no. 2, April 1994, pp. 322-347.

Kamien, Morton I. and Nancy L. Schwartz. *Dynamic Optimization: The Calculus of Variations and Optimal Control in Economics and Management*, 2nd ed. Amsterdam: Elsevier Science, 1991.

Klemperer, Paul. "Competition when Consumers Have Switching Costs: An Overview with Applications to Industrial Organization, Macroeconomics, and International Trade." *Review of Economic Studies* 62, 1995, pp. 515-539.

Klepper, Steven. "Entry, Exit, Growth, and Innovation over the Product Life Cycle." *American Economic Review* 86 no. 3, 1996, pp. 562-583.

Klepper, Steven, and Kenneth L. Simons. "Technological Extinctions of Industrial Firms: An Enquiry into their Nature and Causes." *Industrial and Corporate Change* 6, 1997, pp. 379-460.

Klepper, Steven, and Kenneth L. Simons. "The Making of an Oligopoly: Firm Survival and Technological Change in the Evolution of the U.S. Tire Industry." *Journal of Political Economy* 108, 2000a, pp. 728-760.

Klepper, Steven, and Kenneth L. Simons. "Industry Shakeouts and Technological Change," *International Journal of Industrial Organization* 23 (1-2), February 2005, pp. 23-43.

Liebowitz, S. J. and Stephen E. Margolis. "Path-Dependence, Lock-In, and History." *Journal of Law, Economics, and Organization*, 1995, pp. 205-226.

Nelson, Richard R. and Sidney G. Winter. "Forces Generating and Limiting Concentration under Schumpeterian Competition." *Bell Journal of Economics* 9 no. 2, 1978, pp. 524-548.

Penrose, Edith. *The Theory of the Growth of the Firm*. Oxford: Basil Blackwell, 1959.

Schmalensee, Richard. "Inter-Industry Studies of Structure and Performance." In Richard Schmalensee and Robert D. Willig, eds., *Handbook of Industrial Organization, Volume 2*, Amsterdam: Elsevier, 1989, pp. 951-1009.

Seierstad, Atle and Knut Sydsæter. *Optimal Control Theory with Economic Applications.* Amsterdam: Elsevier Science, 1987.

Shaked, Avner and John Sutton. "Product Differentiation and Industrial Structure." *Journal of Industrial Economics* 36 no. 2, 1987, pp. 131-146.

Simons, Kenneth L. "Product Market Characteristics and the Industry Life Cycle." Manuscript, Rensselaer Polytechnic Institute, 2005.

Sutton, John. *Sunk Costs and Market Structure: Price Competition, Advertising, and the Evolution of Concentration*. Cambridge, Mass.: MIT Press, 1991.

Sutton, John. *Technology and Market Structure: Theory and History*. Cambridge, Mass.: MIT Press, 1998.

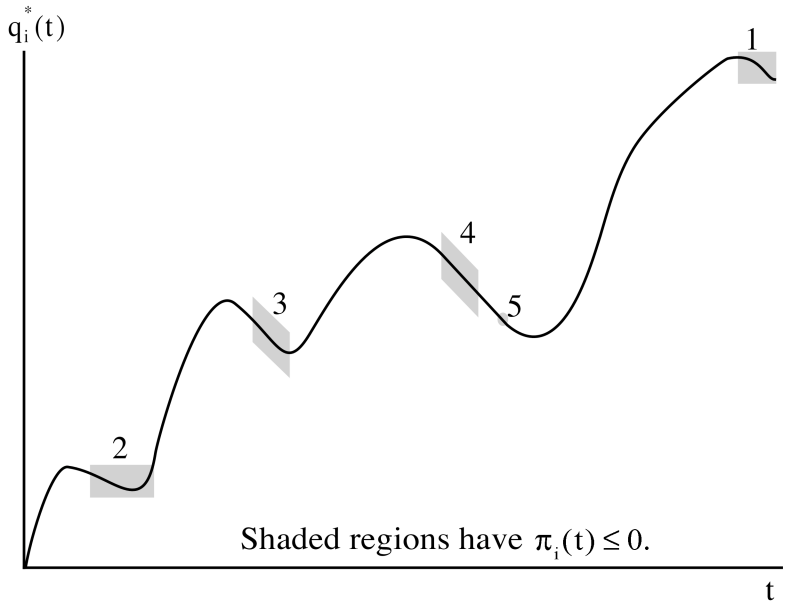

Figure 1. Method of Proof for Lemma 6

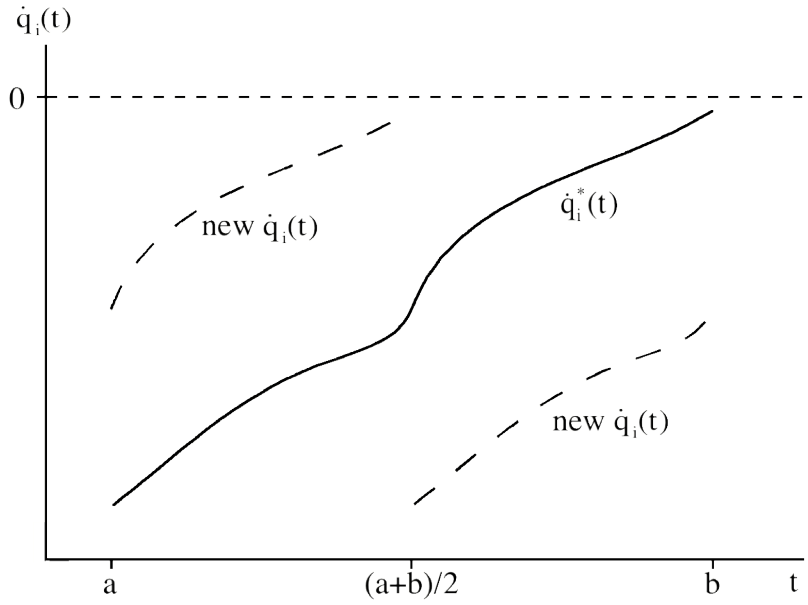

Figure 2. Method of Proof for Lemma 6 Step 4

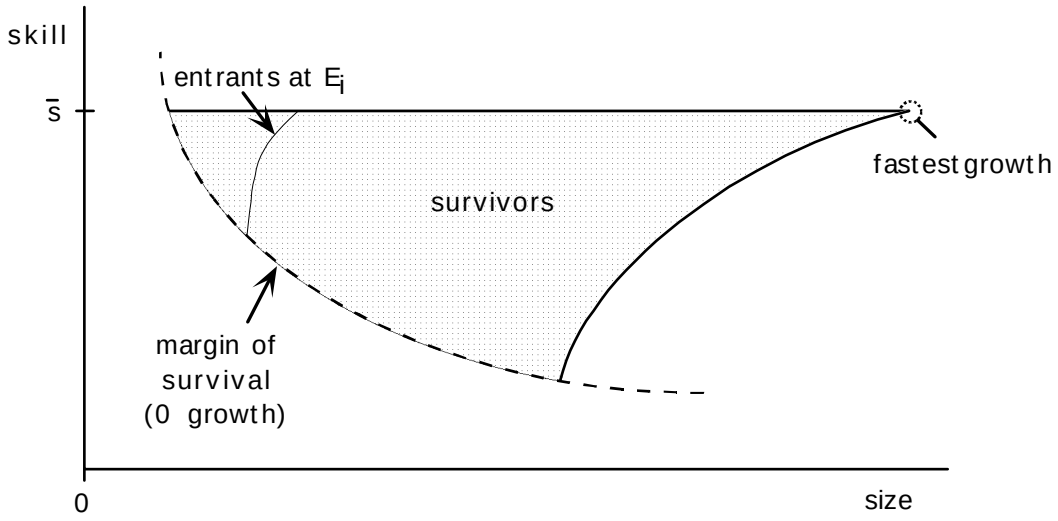

Figure 3. Survivors' Size and Skill at a Time τ After Entry Ceases