Input choices and rate-of-return regulation: an overview of the discussion

William J. Baumol
Professor of Economics
Princeton University

and

Alvin K. Klevorick
Associate Professor of Economics
Yale University

This article reviews the substance of the literature stemming from the Averch-Johnson model of the firm under rate-of-return regulation. It examines a number of propositions, among them the following: (1) The profit-maximizing firm under rate-of-return regulation will tend to use a capital-labor ratio greater than that which minimizes cost for its output level; (2) The profit-maximizing firm under regulatory constraint will use a capital-labor ratio and produce an output greater than it would in the absence of regulation; (3) The closer the "fair rate of return" is to the true cost of capital, the greater the quantity of capital the firm will want to use; and (4) The sales (total revenue)-maximizing firm under regulatory constraint will use a labor-capital ratio greater than it would when unconstrained. The second of these propositions, which has been widely taken to have been proved by Averch and Johnson, is shown to be false. The paradoxical assertion in the third proposition is explained and its regulatory implications discussed. Two models involving regulatory lag are described along with their implications for policy. Finally, some evaluative comments are offered on the entire issue of what has come to be known as the "A-J effect" and its importance in regulatory economics.

The imposition of an artificial constraint on a decision process that would otherwise be optimal will certainly lead to no improvement in outcome and will generally make things worse. In particular, a regulatory constraint on the firm's rate of return, if it affects input proportions at all, may be expected to reduce their efficiency. This issue was explored explicitly by Harvey Averch and L. L. Johnson.
in a pathbreaking article in 1962 [11], using a model similar to one constructed almost simultaneously in a paper written independently by S. H. Weilisiz [11]. These authors showed within a static equilibrium framework that, given a number of conditions to be discussed presently, rate-of-return regulation will indeed modify the capital-labor ratio utilized by the profit-maximizing firm. Specifically, their model led to a demonstration that under the conditions postulated the firm will find it profitable to employ more capital relative to labor than is consistent with minimization of cost for the quantity of output produced.

These results have led to a considerable discussion, technical and informal, in terms of abstract general principles and concrete policy implications. The writings on the subject have disputed the rigor of the original proofs and supplied alternative derivations; they have raised questions about the assumptions and expressed doubts about the applicability or the importance of the results in practice.

Without surveying the literature item by item, this paper seeks to review the substance of the discussion. It first restates the relevant theorems and assertions and provides informal and formal proofs. It then turns to some related propositions that have been developed and discusses some work currently in progress in the area. Finally, it offers, for whatever they may be worth, our views of the upshot of the discussion. In sum, we conclude that the analytical approach employed by the original authors is a very fruitful one. Indeed, their piece represents a breakthrough in the theory of regulation, constituting a major contribution to the rigorous analysis of the principles of regulation of industry, a field which has recently blossomed. Moreover we agree that there are dangers to economic efficiency in the rate-of-return approach to rate regulation. Nevertheless, we will suggest that the phenomenon that emerges from the A-J theorem may not be of very great significance in practice. It is at least plausible that other potential sources of difficulty in the regulatory process dwarf the consequences of the distortion in the capital-labor ratio that the model predicts.

The original Averch-Johnson analysis provides a simple and comprehensible depiction of a regulatory process and its relation to the behavior of the firm. The company is taken to produce one output and to use two inputs, labor and capital, each of which is available to it in unlimited quantities at a fixed price per unit. The firm's decisions are affected by the regulator in only one way: it is permitted to earn no more than some fixed proportion of the value of its capital—the regulatory “fair rate of return” on its rate base. In all other respects management is permitted to pursue its objective, maximization of profit, exactly as it could in the absence of regulation. Finally, it is assumed that the rate of return permitted to the firm by regulation is less than the return it would obtain if it were able to maximize its profits, but is at least as great as the rate of return on capital. These assumptions can be brought out more

1. The A-J model of the regulatory process

---

1 While the Weilisz model deals with the same sort of regulated firm, it concerns itself primarily with the effects of regulation on peak-load pricing rather than on output usage. The Weilisz discussion will, therefore, not be examined in this article.
explicitly with the aid of mathematical notation, which largely follows that in the original A-J article.

Let

\[ \pi = \text{the firm's total profit}, \]
\[ z = \text{the physical quantity of the firm's output}, \]
\[ x_1 = \text{the quantity of capital in the firm's base},^2 \]
\[ x_2 = \text{the quantity of labor input}, \]
\[ r_1 = \text{the unit cost of capital = the opportunity cost of resources tied up in plant and equipment}, \]
\[ r_2 = \text{the unit cost of labor}, \]
\[ s = r_1 + v \text{ be the rate of return permitted by regulation, } v \geq 0, \]
\[ z = z(x_1, x_2) = \text{the production function, } \frac{\partial z}{\partial x_1} > 0, \]
\[ \frac{\partial z}{\partial x_2} > 0 \text{ and } z(0, x_2) = z(x_1, 0) = 0, \]
\[ p = p(z) = \text{the inverse demand function.} \]

Then the company is assumed to select the values of \( z, x_1 \) and \( x_2 \) that maximize total profit

\[ \pi = px_1 - r_1 x_1 - r_2 x_2 \tag{1} \]

subject to the regulatory constraint

\[ \frac{px_1 - r_2 x_2}{x_1} \leq r_1 + v = s . \tag{2} \]

Specifically, since \( r_1 + v \) is taken to be less than the profit-maximizing rate of return, it is assumed by A & J that the constraint (2) is an equation rather than an inequality.\(^3\)

Notice the very rudimentary regulatory process that is encompassed in this model. First, it is assumed that the firm can decide for itself on any price-output combination it wishes, so long as the rate-of-return constraint is satisfied. Second, it will be assumed at least for part of the discussion (following most of the literature) that the rate-of-return constraint is satisfied as an equality at all times and that there is no regulatory lag. If the regulatory commission decrees that net returns will not exceed \( s \) percent of the rate base, then they will equal \( s \) percent precisely, no more and no less, even as a temporary matter. Clearly, this is not quite the way the regulatory process works, though the model's simplification of the facts is not unreasonable as a first approximation. We shall see later what analytical difficulties one gets into in seeking to represent the regulatory process more closely.

The firm in this model is also rather rudimentary, its production process involving only two inputs and one output, and its objective being total profit maximization pure and simple. However, the choice of objective aside, no one seems to have suggested that any significant loss of generality results. The analysis also adopts several common assumptions as a matter of mathematical convenience: all relevant functions are assumed twice differentiable, and the appro-

---

\(^2\) Because of depreciation the value of such resources will normally be less than the acquisition cost of these capital assets. For simplicity we will follow Averch and Johnson in assuming depreciation to be zero and the acquisition cost of capital to be unity.

\(^3\) If capital were to depreciate, the rate of return in equation (2) would have to be expressed, not in terms of the acquisition cost of the capital, but in terms of its depreciated value. That is, \( x_1 \) would have to be the depreciated value of the company's capital, and current depreciation would appear as an operating cost. As just stated, it is convenient to assume this complication away.
appropriate second-order maximum (concavity-convexity) conditions are taken to hold throughout. Some of these will be discussed explicitly later.

II Under the circumstances postulated it is easy to prove that the capital-labor ratio chosen by the firm will, in general, be different from the one that minimizes the cost of whatever output is decided upon. It is only slightly more difficult to show that the resulting input proportions will be more capital-intensive than efficiency requires. There has been some discussion of the validity of the argument used by A & J in arriving at this second result. As a matter of fact, as we shall see presently, it may not even be the result they had in mind when they wrote their paper. For such reasons, we find it useful to divide the hypotheses explicit and implicit in the discussion into several different propositions. We begin with

**Proposition 1.** The firm described in the model (1) and (2) will adopt input proportions different from those that minimize the cost of the final output level, $z^*$. 

**Proof:** Minimization of the cost of the output selected requires minimization of $r_1 x_1 + r_2 x_2$ subject to $z(x_1, x_2) = z^*$, which yields the Lagrangian

$$r_1 x_1 + r_2 x_2 + \mu (z^* - z(x_1, x_2))$$

and this leads to the first-order conditions

$$r_1 - \mu z_1 = 0, \quad r_2 - \mu z_2 = 0$$

where $z_i$ denotes $\partial z / \partial x_i$, the marginal product of $x_i (i = 1, 2)$ or $r_i / r_2 = z_1 / z_2$. (3)

Turning next to the regulated firm, its Lagrangian is, by (1) and (2) (treating the latter as an equality until otherwise noted),

$$L = p z - r_1 x_1 - r_2 x_2 - \lambda (p z - s x_1 - r_2 x_2).$$ (4)

We now obtain the first-order conditions

$$\begin{align*}
(1 - \lambda) MR_z z_1 - r_1 &= - \lambda s, \text{ i.e., } (1 - \lambda) MR_z z_1 \\
&= (1 - \lambda) r_1 = \lambda (r_1 - s) \quad (5a) \\text{ where } \quad (1 - \lambda) MR_z z_2 - (1 - \lambda) r_2 = 0 \quad (5b)
\end{align*}$$

We will follow the literature in assuming through most of the discussion that

$$s > r_1 \quad (6)$$

(the “fair rate of return” exceeds the cost of capital). It then follows

---

4 Averch and Johnson utilize the Kuhn-Tucker conditions at this point because the regulatory constraint (2) is an inequality. However, they always assume this constraint is effective, so that they treat it as an equation throughout their discussion. As Takayama [10] points out, it is consequently unnecessary to use the Kuhn-Tucker conditions.
from (5a) that $\lambda \neq 1$. For if $\lambda = 1$ (5a) becomes simply $r_1 = s$, thereby contradicting (6). Thus we may divide equation (5a) by (5b), obtaining, after some rearrangement,

$$z_1/z_2 = r_1/r_2 - \lambda(s - r_1)/(1 - \lambda)r_2.$$  \hfill (7)

Thus, if we can show that the last term in (7) is not equal to zero, it will follow that requirement (3) for the least-cost input combination is violated by our regulated firm. Because we have $s > r_1$ and $\lambda \neq 1$ as was just shown, we need only to show that $\lambda \neq 0$. This follows from comparison of the Lagrangian (4) with the original objective function (1). For if $\lambda = 0$ the two expressions are identical, so that the constrained maximization problem will yield a profit level identical with that which can be achieved by a firm unconstrained by regulation. Since we have assumed that the rate-of-return constraint does reduce company profits, it follows that we cannot have $\lambda = 0$. This completes the proof of the first A-J theorem.

A number of the published discussions leap from this result to one of the following two conclusions:

(Aleged) Proposition 2. The capital-labor ratio of the regulated firm will be larger than that of the unconstrained profit-maximizing monopolist.

Proposition 3. The capital-labor ratio of the regulated firm will be larger than the one that minimizes costs for the output level that it elects to produce.

It is not quite clear which of these A & J had in mind when they wrote their article. Certainly their discussion on this point is ambiguous (compare their Figure 1 with their discussion on p. 1053, top, and 1056, bottom).

As we shall see, propostion 2 is simply false, as a generalization, while proposition 3 is correct under the circumstances postulated although details of the proof require some care. We shall, in section 6, provide a counterexample to proposition 2 and discuss its problems in general terms. For the moment we shall confine our discussion to proposition 3, which is the more significant of the two from the point of view of efficiency of resource utilization. We will therefore present in detail a proof of the result. A critical step in the argument is the following lemma, which we will also utilize later in this article.

Lemma. If the firm maximizes its total profit (1) subject to the regulatory (inequality) constraint (2), in which $s > r_1$, and if, in addition, the regulatory constraint is binding and $x_1 > 0$ and $x_2 > 0$, then we must have $0 < \lambda < 1$.

Since the discussion of this lemma in the literature is somewhat disjointed and impressionistic (see Takayama’s criticism of A & J on this score), a more systematic derivation seems appropriate. Maximization of (1) subject to (2) yields as its first-order (Kuhn-Tucker) conditions with $x_1 > 0$, $x_2 > 0$, equations (5a), (5b), and

$$pz - sx_1 - r_2x_2 \leq 0$$  \hfill (5c)

$$\lambda(pz - sx_1 - r_2x_2) = 0$$  \hfill (5d)
\[ \lambda \geq 0. \quad (5e) \]

Writing \( R_1 \) as an abbreviation for \( MR_{x_1} \), the marginal revenue product of \( x_1 \), and using similar notation for \( x_2 \), we obtain by solving\(^5\) (5a) for \( \lambda \),

\[ \lambda = \frac{(r_1 - R_1)}{(s - R_1)}. \quad (8) \]

The unconstrained profit maximum is given by the values of \( x_1, x_2 \) satisfying

\[ R_1 = r_1, \quad R_2 = r_2. \]

By (5b), the second of these conditions is satisfied also by the regulated firm. Hence, since the regulatory constraint is assumed to be effective, i.e., it prevents the firm from maximizing profits, we must have

\[ R_1 \neq r_1 \]

and consequently, by (8), \( \lambda \neq 0 \). Therefore by (5e)

\[ \lambda > 0. \]

We have already seen, in the discussion immediately following (6), that \( \lambda \neq 1 \). To prove that \( \lambda < 1 \) we utilize the second-order conditions for the constrained profit maximum, which require a positive value for the bordered Hessian, i.e.,

\[
H = \begin{vmatrix}
(1 - \lambda)R_{11} & (1 - \lambda)R_{12} & s - R_1 \\
(1 - \lambda)R_{21} & (1 - \lambda)R_{22} & r_2 - R_2 \\
s - R_1 & r_2 - R_2 & 0
\end{vmatrix} > 0
\]

where \( R_{ij} \) denotes \( \partial^2 p_z / \partial x_i \partial x_j \). However, by (5b) \( r_1 - R_1 = 0 \). Therefore the preceding condition becomes (on expanding the determinant in terms of its last column),

\[ H = - (s - R_1)(1 - \lambda)R_{22} > 0. \]

Diminishing revenue returns to labor [concavity of \( R(x_1, x_2) \)] gives us \( R_{22} < 0 \) so that we obtain immediately

\[ (1 - \lambda) > 0 \]

as was to be shown.

It follows, incidentally, that \( R_1 < r_1 < s \), i.e., that for the regulated firm in equilibrium the marginal revenue product of capital will be less than \( r_1 \), the cost of capital. This follows from (8), which indicates that if \( r_1 < R_1 < s \) then \( \lambda < 0 \), while if \( r_1 < s < R_1 \) then \( \lambda > 1 \).

We now return to the main algebraic argument adduced in support of proposition 3. We see at once from the preceding lemma that the last term in (7) must be negative since \( 0 < \lambda < 1 \) and since, by (6), \( s > r_1 \). From this one obtains the required result. Having shown by (7) that

\[ z_1/z_2 < r_1/r_2, \quad (9) \]

comparison with (3) demonstrates that the relative use of capital must, under (9), exceed the ideal proportion corresponding to (3). For \( z_1/z_2 \), the marginal rate of substitution of capital for labor, is now below the ratio of input prices \( r_1/r_2 \). With diminishing marginal revenue returns to labor, the excess of \( z_1/z_2 \) over that of \( r_1/r_2 \) indicates that the efficient use of the capital stock for production is not obtained.

---

\(^5\) The division by \( s - R_1 \) is legitimate because if \( s = R_1 = MR_{x_1} \) we must have, by (5a), \( s = r_1 \), and this is ruled out by (6).
rate of substitution of capital for labor for a fixed output level, this can occur only as the result of a relative increase in the use of capital.

We deal next with the remaining contention of the A-J paper:

(ALLEGED) PROPOSITION 4. The regulated firm will produce an output larger than that which maximizes profits.

Strangely, one looks in vain for any attempt at a formal proof of this assertion. A & J seem to arrive at their result through a simple heuristic argument: From (5a) or (9) they conclude that the regulated firm will purchase more capital than the amount that maximizes profit [actually this does not follow directly from the equations cited, though, as will be shown later (proposition 5), the assertion about the capital stock is correct]. Similarly, equation (5b) can be rewritten simply as $MR_{z_3} = r_3$, asserting that the marginal revenue product of labor $(x_3)$ will be equal to the wage level $(r_3)$. This seems to have suggested that the profit-maximizing quantity of labor will be hired. With that quantity of labor and with the quantity of capital greater than the amount that maximizes profit, it follows that output must be greater than the profit-maximizing value of $z$.

As already noted, it is not clear that this is the argument that leads A & J to their allegation about the firm's output level. In any event, the argument is simply incorrect. Condition (5b) does not require the quantity of labor hired by the regulated firm to be the same as the amount that would be employed under pure profit maximization. The quantity of labor employed by the firm will still depend on the quantity of capital it uses and the nature of the production function. As Takayama [10] shows, rate-of-return regulation can in fact lead to a decrease in the quantity of labor hired by the firm, a point to which we will return presently. Indeed we will show later that, as a generalization, proposition 4 simply is not valid.

The invalidity of proposition 4 may perhaps be considered unfortunate. If it were true it might have been claimed as a virtue of the regulatory process. For generally the monopolistic firm's output is taken to be smaller than the level that is socially optimal. If regulation of rate of return were to guarantee an increase in output above the monopolistic level it might have been considered to be most desirable. Indeed, some authors have defended fair-rate-of-return regulation on the ground that it will have a beneficial effect on output. For example, Alfred E. Kahn [6] recognizes explicitly the input inefficiency such regulation can engender. But he argues, by inducing the regulated firm to overinvest in capacity and hence to increase its output, fair-rate-of-return regulation helps to reduce—if not overcome—the monopolist's tendency to underinvest and restrict output. The fact that proposition 4 is not valid as a general statement weakens somewhat the case of Kahn and others who take a similar position.

3. The geometry of the A-J propositions

WILLIAM J. BAUMOL AND ALVIN K. KLEVORICK

As it turns out, a good deal of illumination can be contributed by translation of the model's relationships into geometric terms. Two closely related graphic interpretations have been provided independently by Zajac [14] and by Klevorick and Baumol [8]. The discussion here will employ elements of both analyses.
Figure 1 is a three-dimensional diagram depicting on its axes the values of $x_1$, $x_2$, and total profit. It also shows several planes, all of them hinged on the $x_2$ axis and hence satisfying the equation

$$
\pi = vx_1
$$

for different values of $v$. Any such plane represents the rate-of-return constraint (2) for a particular value of $v$. The interpretation of the left-hand side of (10) as a profit function follows directly from (1) and (2). Rewriting our regulatory constraint (2) as an equality, as we are now treating it, and substituting this into the original profit function (1) we obtain at once

$$
\pi = (p_2 - r_2x_2) - r_1x_1 = sx_1 - r_1x_1 = vx_1,
$$

Figure 2
which is an expanded restatement of (10). We see from (11) that the slope of any of the cross sections of a constraint plane taken parallel to the \((x_1, \pi)\) plane represents \(s - r_1\), the excess of the regulatory "fair rate of return" over the cost of capital.

In Figure 2 the firm's profit function has been included in the diagram as the shaded hill-shaped figure. The heavy broken curve is the intersection of the profit surface and the regulatory constraint plane. Its projection on the \(x_1x_2\) plane represents all combinations of the two inputs which just satisfy the rate-of-return constraint requirement. The heavy curve in Figure 3 depicts that projection on the \(x_1x_2\) plane. The shaded region inside the constraint curve (the maximal-rate-of-return curve) corresponds to input combinations that yield rates of return higher than the maximum permitted by regulation. The heavy constraint curve is an iso-rate-of-return locus; i.e., points on that curve represent input combinations all offering the same rate of return on capital. Three other iso-rate-of-return loci, all of them offering higher rates of return than permitted by the constraint, are also shown. Point \(\pi_{\text{max}}\) corresponds to the highest point on the profit surface: it is the point the profit-maximizing firm would select in the absence of regulation.\(^6\) The point \(\pi_{\text{reg}}\) is the input combination that maximizes profit given the regulatory constraint.\(^7\) A moment's thought shows that \(\pi_{\text{reg}}\) must be the rightmost point on the regulatory constraint curve since this gets the firm to the highest total profit point on the constraint plane in Figure 2. Looked at another way, the rightmost point maximizes total profit consistent with the constraint because all points on the constraint curve yield the same net rate of return, \(r\), and the rightmost point involves the largest rate base, \(x_1\), and hence the largest value of \(\pi = nx_1\).

\(^6\) Note that point \(\pi_{\text{max}}\) which yields maximum total profits does not offer the maximal rate of return. For, as the diagram shows, there are iso-rate-of-return loci inside the curve through \(\pi_{\text{max}}\) whose rates of return are higher than that through \(\pi_{\text{max}}\).

\(^7\) While in this diagram the output at \(\pi_{\text{reg}}\) is greater than that at \(\pi_{\text{max}}\) as proposition 4 claims is generally true, we shall show presently that this result does not always hold.
The broken curve in Figure 3 is the locus of efficient combinations of labor and capital. For any given output level the least-cost input combination will be given by some point on this curve, \( \pi_{\text{max}}E \).

Proposition 3 corresponds to the assertion that the constrained maximum point lies below (to the right of) the efficient locus, \( \pi_{\text{max}}E \). For if this is true then \( QQ' \), the iso-product curve through point \( \pi_{\text{reg}} \), will cross the efficiency locus at some point \( K \) that lies above and to the left of \( \pi_{\text{reg}} \) (since iso-product curves must have a negative slope in capital-labor space). \( K \) is, by definition of the efficient locus, the least expensive input combination for the production of the output level corresponding to \( QQ' \). That is, \( K \) produces the same output as \( \pi_{\text{reg}} \) but at lower cost, and \( K \) clearly involves more labor and less capital than \( \pi_{\text{reg}} \). Thus, if we can prove that \( \pi_{\text{reg}} \) lies below and to the right of the efficiency locus, it follows that the regulated firm in our model will utilize an inefficiently high capital-labor ratio to produce the output level it selects, as proposition 3 requires.

It therefore only remains for us to show that point \( \pi_{\text{reg}} \) does not lie on or above the efficiency locus \( \pi_{\text{max}}E \). To prove this, assume the contrary, as in the case depicted in Figure 4, where \( \pi_{\text{reg}} \) lies above the efficient locus. Let point \( A \) be the intersection point of the efficient curve with \( QQ' \), the iso-product locus through \( \pi_{\text{reg}} \) and let \( B \) be the intersection of the efficient locus with the constraint curve.

Obviously, \( A \) is more profitable than \( \pi_{\text{reg}} \) since it yields the same output at a lower production cost. Moreover, by the concavity conditions implicit in the argument, profit increases monotonically as one moves along the efficiency locus toward \( \pi_{\text{max}} \). Thus \( B \) must be at least as profitable as \( A \). Hence \( B \) must be more profitable than \( \pi_{\text{reg}} \), contradicting the definition of \( \pi_{\text{reg}} \) as the most profitable point on the constraint locus.

This completes the Zajac proof of proposition 3.

There is an alternative geometric derivation of the A-J result that is also illuminating. Figure 5a depicts the firm's iso-profit curves—curves of constant total profit, the contour lines for the total profit surface that was represented in Figure 2. The concavity of the profit
surface assures us that each such curve is the boundary of a convex region.

We shall use this diagram to describe explicitly the location of the firm's efficient locus, \( E'E' \), and of its expansion path \( RR' \), showing the combinations of inputs that will be hired by the regulated firm with different values of \( s \), the permitted rate of return. That is, \( RR' \) is the locus of all equilibrium points, \( \pi_{eq} \), for our firm. If we can show that everywhere to the right of the maximum profit point, \( \pi_{max} \), the locus \( RR' \) lies below and to the right of the efficient locus, as it does in Figures 5a and 5b, then the A-J result must follow. That is, under rate-of-return regulation, in producing a given quantity of output the firm must utilize a capital-labor ratio larger than that required for maximal efficiency. For since (Figure 5b) the iso-product curve \( QQ' \) has a negative slope, its intersection point, \( B \), with the lower expansion path locus, \( RR \), must lie below and to the right of \( A \), the efficient point for that output.

We prove our assertion about the relative positions of \( RR' \) and \( EE' \) as follows: Curve \( EE' \) is the locus of tangency points, \( T_i \), between the iso-profit curves and the parallel negatively sloped price lines \( P_iP_i' \). Hence curve \( EE' \) must intersect the iso-profit curves in their negatively sloping segments such as \( H_1V_1 \) in Figure 5a.

However, the expansion path, \( RR' \), is the locus of right-hand vertical points on the iso-profit curves. Our result follows at once from the convexity properties of these curves: A tangency point, \( T_i \),

---

\( ^8 \) As is shown in proposition 5, below, when the rate of return constant is effective, the firm will always use a quantity of capital greater than that at \( \pi_{eq} \). Hence, any solution point must lie to the right of \( \pi_{max} \) in the diagram. This also follows directly from Figure 3 where we see that \( \pi_{eq} \) is the constrained maximum point, must always lie to the right of \( \pi_{max} \) since \( \pi_{eq} \) is the rightmost point on the constant curve.

\( ^9 \) The slope of the price lines is, of course, equal in absolute value to the ratio of input prices, \( r_1/r_2 \). The slope of an iso-profit curve (see the following note) is \( -\frac{v_1}{v_2} \) where \( \tau_i = \frac{\partial v_i}{\partial x_i} \). Hence, tangency between an iso-profit curve and a price line requires

\[ \frac{r_1}{r_2} = \frac{\pi_{eq}}{\pi_{max}} = \frac{(MR_i, -\tau_1)(MR_j, -\tau_2)}{v_1 - v_2} = \frac{\tau_i MR_i, -\tau_2}{\tau_1 MR_i, -\tau_1} \]

\( \tau_1 / \tau_2 = r_1 / r_2 \), which is the same as the necessary condition for efficiency (3).

\( ^{10} \) Proof. The iso-profit curve is given by \( v = \text{constant} \), from which we deduce \( \pi dx_1 + v dx_2 = 0 \), that is, the slope of the iso-profit curve is \( dx_1 / dx_2 = -v_1 / v_2 \). But from (5b) we obtain \( \tau_1 MR_i x_1 - v_1 = 0 \). Hence, equilibrium of the regulated firm thus requires the iso-profit curve to be vertical.

An argument that is more geometric in spirit is also possible. We have already noted that equilibrium of the regulated firm requires it to be on the vertical (right-hand) point of an iso rate-of-return curve \( v = \text{constant} \), since at this point \( x_2 \), and hence \( v \), will be at a maximum. The slope
on the negatively sloping portion of any one of these curves on its rightmost end must lie above and to the left of the corresponding vertical (right-hand) point, \( V_r \).

A & J emphasize that their results depend on the assumption that the regulatory "fair rate of return" exceeds the cost of capital. This assertion is not very easy to test empirically, since the cost of capital presumably varies from firm to firm, depending at least in part on the degree of risk incurred by anyone who provides capital to it, and we do not really know how to measure that cost of capital with any high degree of precision. The proceedings of the regulatory commissions that determine the firm’s fair rates of return certainly do have the appearance of a search for a figure approximating the cost of capital. They argue in terms of the return necessary to enable the firm to attract the capital it needs, and that is presumably what is meant by the (opportunity) cost of capital. In other words, it is at least arguable that the regulator often attempts to set the "fair rate of return" at a level as close to the cost of capital as he can determine.

For this reason, if not as a purely analytic matter, it is of interest to examine what happens to the A-J propositions as the regulated rate of return, \( r_1 \), approaches the cost of capital, \( r_1 \), and what occurs when the two coincide. As already implied, A & J themselves point out that their result breaks down when \( s = r_1 \). However the nature of this breakdown is a bit surprising. Our intuition might suggest that once capital no longer receives a return that may be termed "excessive," the firm will find it unprofitable to use an inefficiently large quantity of capital and it will therefore be motivated to utilize the efficient capital-labor ratio. As a matter of fact, when \( s = r_1 \) the firm has no motivation to select this ratio or any other particular ratio. Any capital-labor ratio becomes as profitable to the regulated firm as any other.

This result can be demonstrated with the aid of the first-order maximum conditions (2), (5a), and (5b). When \( r_1 = s \) we see that \( \lambda = 1 \) satisfies (5a) and (5b). But when \( \lambda = 1 \) the firm's capital-labor ratio is indeterminate because the only first-order condition that continues to be effective is the original profit constraint (2). Any values of \( x_1 \) and \( x_2 \) that satisfy the regulatory constraint must, therefore, satisfy all of the first-order maximum conditions, and any such pair of values is as profitable to the regulated firm as any other.

Geometrically we can see more clearly why the A-J results collapse in this peculiar way. Returning to Figure 1, we recall from (10) and (11) that the slope of one of the constraint planes taken parallel to the \( x_1, x_2 \) plane is \( v = s - r_1 \). Hence when \( s = r_1 \) the plane is flat—regulation offers the firm no opportunity to increase its total profit either through an increase or a decrease in the quantity of \( x_1 \). The constraint plane is now perfectly horizontal and, in fact, it coincides with the floor of the diagram—the \( x_1, x_2 \) plane. Regulation of the iso-rate-of-return curve is obtained by total differentiation thus:

\[
d(s/x_1) = \left( \frac{x_1}{x_1} \right) dx_1 + \left( \frac{x_2}{x_1} \right) dx_2 = 0
\]

so that

\[
dx_1/x_1 = -(x_1 - x_1/x_1) dx_2.
\]

At a vertical point on this curve we must therefore have \( x_1 = 0 \), which is precisely the condition for the iso-total profit curve to be vertical since, as we have just seen, its slope is \( dx_1/x_1 = v/s_1 \).
now guarantees that the firm receives for its capital exactly what has to be paid for it, no more and no less. That is, profits will equal zero after payment of the “normal” return to capital, \( r_1 \). In this circumstance the firm has no particular motivation to employ an efficient quantity of capital, nor any motivation to avoid doing so. Entrepreneurial choice in this matter just has no effect on its total profits.

However, it is implausible that in such circumstances management will not adopt some other objective, one which is perhaps secondary but which, if profit is not affected, will be used as a basis for company decisions. It can then be shown that if \( s = r_1 \) and the firm seeks to maximize either total physical volume\(^{11}\) or sales (total revenue), \( p_z \), it is likely to be driven to the ideal input combination point for the corresponding constraint locus, point \( W \) in Figure 3. That is, a regulatory restriction on rate of return with \( s = r_1 \), together with a company goal of sales or output maximization, will yield both ideal output and efficiency in use of resources.

The reason this is so is that maximization of \( p_z \) or \( z \) does require the firm to be on the efficiency locus,\(^{12}\) and the firm will therefore be forced to \( W \), the only such point on our constraint curve.

Once we leave the case where \( r_1 \) and \( s \) coincide, matters seem to follow more closely the course one might expect. As A & J point out, if \( s < r_1 \) the firm loses money and it must, in the long run, go out of business. If, on the other hand, \( s > r_1 \), propositions 1 and 3, the basic results, follow directly. But even here things do not behave quite as one might at first anticipate.

It may seem plausible that the degree of inefficiency in the capital-labor ratio will decline as \( s \) is driven closer to \( r_1 \). That is, it may appear that the more closely the regulator succeeds in pushing the regulatory ceiling return, \( s \), to the cost of capital, the less motivation the firm will have to use inefficiently large quantities of capital. In fact, this turns out to be false. On the contrary, one can show that, at least for some classes of production function, as \( s \) approaches \( r_1 \) the capital-labor ratio will depart further and further from its efficient value! More generally, one can show, following a conjecture of F. M. Scherer, that as \( s \) approaches \( r_1 \) the constrained profit-maximizing quantity of capital will always increase.

This result was proven geometrically by Fred M. Westfield,\(^{13}\) but it was left to Takayama to provide the first rigorous mathematical derivation of the proposition. Curiously though, neither Westfield nor Takayama quite realized the importance of the proposition for regulatory policy. We shall return to its implications after the proof is completed.

Takayama’s argument proceeds as follows. Let \( r_m \) be the rate of return on capital obtained at the (unconstrained) profit-maximizing

\(^{11}\) The idea of maximization of physical volume was suggested to us by Milton Kafatos

\(^{12}\) This can be shown as follows: Since the firm cannot survive with negative profits, the regulatory constraint must be in our present case satisfied as an equality, i.e., we must have \( s = p_z - r_1 = r_1 \). Maximizing \( z \) subject to this constraint we obtain the Lagrangean conditions

\[

\begin{align*}
z_1 + \lambda MR \eta_1 - \lambda \eta_1 &= 0, \quad \text{i.e., } z_1 (1 + \lambda MR) = \lambda \eta_1, \\
z_1 + \lambda MR \eta_2 - \lambda \eta_2 &= 0, \quad \text{i.e., } z_1 (1 + \lambda MR) = \lambda \eta_2, \\
\end{align*}

\]

and straightforward division of one relationship by the other shows that the efficiency condition \( z_1/z_2 = r_1/r_2 \) must be satisfied. The proof for the sales maximization case where the maximand is \( p_z \) rather than \( z \) follows the same argument step by step.

\(^{13}\) In (12), Westfield’s argument employed a geometric apparatus totally different from the one presented in section 3. For example, his diagram is drawn in capital-profits space rather than capital-labor space. See (12), pp. 429–32

WILLIAM J. BAUMOL AND ALVIN K. KLEVORICK

174 / ALVIN K KLEVORICK
input-output combination. Then we have \( r_m > s > r_1 \). For this range of values of \( s \) we can conduct a comparative statics analysis of the value of \( dx_1/ds \). We will prove

**Proposition 5.** For the firm that seeks to maximize total profit (1) subject to the regulatory constraint (2) we have \( dx_1/ds < 0 \) for \( r_1 < s < r_m \).

In other words, the greater the difference between the regulatory fair rate of return and the cost of capital (since we are increasing \( s \) holding \( r_1 \) constant), the smaller will be the value of \( x_1 \), the firm's use of capital.

This result is surely counterintuitive, at least at first blush.

The proof is not difficult. It follows from total differentiation of the first-order conditions (5a), (5b), and (2), utilizing the standard comparative statics procedure to determine how the values of \( x_1 \) and \( x_2 \) respond to a change in the value of \( s \), holding \( r_1 \) constant. For simplicity, let us again rewrite total revenue \( p_z \) as \( R(x_1, x_2) \) so that, e.g., \( R_1 = \partial R/\partial x_1 \) represents the marginal revenue product of \( x_1 \). Then the first-order conditions (5a), (5b) and (2) become (treating the regulatory constraint as an equality)\(^{14}\):

\[
\begin{align*}
R_1 - r_1 &= (r_1 - s)\lambda/(1 - \lambda) \quad (12) \\
R_2 - r_2 &= 0 \quad (13) \\
R - r_2x_2 &= sx_1.
\end{align*}
\]

Differentiating the last of these conditions totally, one obtains

\[
(R_1 - s)dx_1 + (R_2 - r_2)dx_2 = x_1ds. \quad (14)
\]

By (13) the second term in this equation is zero. Moreover, by (12)

\[
R_1 - s = r_1 - s + (r_1 - s)\lambda/(1 - \lambda) = (r_1 - s)/(1 - \lambda). \quad (15)
\]

Thus, substituting this relationship into (14) and after dropping its second term by (13) we have

\[
dx_1/ds = (1 - \lambda)x_1/(r_1 - s) < 0. \quad (16)
\]

This is negative since the lemma of Section 2 gave us \( 0 < \lambda < 1 \) and since \( r_1 < s \) by assumption. Q.E.D.

It is worth examining this result somewhat more carefully. As already indicated, Takayama asserts that this is an alternative proof of the A-J theorem, but that is simply incorrect. Neither proposition 2 nor 3 follows from his argument. True, he has shown that the imposition of the constraint, as it reduces the rate of return from the profit-maximizing level \( r_m \) to the legal ceiling \( s \), does increase the firm's use of \( x_1 \). But provided that the firm simultaneously expands its use of \( x_3 \) sufficiently, there is no need for the resulting capital-labor ratio to be inefficient. Indeed, as the value of \( s \) decreases we would normally expect an efficient value of \( x_4 \) satisfying the regulatory constraint to increase as well. For by increasing \( x_1 \) and \( x_2 \) sufficiently along the efficient path, profit can be driven down to the level required by a lower value of \( s \). In terms of Figure 3, a lower rate-of-return ceiling moves the constraint curve further away from \( \pi_{\text{max}} \), and therefore it moves the profit-maximizing point, \( \pi_{\text{reg}} \), the vertical

\(^{14}\) Recall that since \( r_1 < s \) we know that \( \lambda < 1 \) and since \( s < r_m \) we have \( \lambda > 0 \) (the Lemma of section 2)
point on the constraint curve, further to the right. That is precisely why a decrease in the permitted rate of return necessarily increases $x_1$, the quantity of capital that will be used by the firm. However, the reduction in value of $s$ also leads the efficient point, $W$, on that curve to move further to the right, meaning that a larger use of capital is also required to maintain efficiency. In sum, there is no necessity of inefficiency implicit in the result that the value of $x_1$ is increased by regulation.

This moderately painstaking discussion of the Takayama result was undertaken primarily because of the importance of the proposition for regulatory policy. For proposition 5 at least yields the disquieting result that a better job of setting the fair rate of return close to the cost of capital will not reduce the absolute capital bias and that it need not reduce the inefficiency in the input proportions shown by proposition 3. On the contrary, the result of better estimation of the cost of capital by the regulator may simply be to aggravate that inefficiency if the rise in the use of capital is not matched by an appropriate increase in the quantity of labor utilized.

A concrete example will demonstrate that this can in fact happen, i.e., that as $s$ approaches $r_1$ not only will $x_1$ increase but there can also be an increase in the difference between the profit-maximizing capital-labor ratio $x_1/x_2$ and the efficient value of that ratio.

For this purpose let the total revenue function be

$$R(x_1, x_2) = a \ln x_1 + b \ln x_2$$

where $a > 0$, $b > 0$ so that $R(x_1, x_2)$ is strictly concave. Then as our marginal revenue products we have

$$R_1 = a/x_1 \quad R_2 = b/x_2,$$

By (3) the efficient path is given by $R_1/R_2 = r_1/r_2$, i.e., by

$$r_1 x_1 = r_2 x_2,$$

which is clearly a straight line through the origin in $(x_1, x_2)$ space. Thus, in this case the efficient path involves a constant capital-labor ratio.

On the other hand, we know by (16) that $dx_1/ds < 0$ while, as we will see presently, from (18), $dx_2/ds$ must in this case equal zero since $R_{x2} = 0$. Thus we see that as $s$ moves closer and closer to $r_1$ the constrained profit-maximizing $x_1/x_2$ ratio must always increase, and hence it must constantly depart farther from the constant proportion between $x_1$ and $x_2$ that is required for efficiency.  

---

5. The demand for labor and the profit-maximizing output level

One may be tempted to infer, carelessly, that since by (5b) $R_1 = r_1$, the regulated firm will employ the optimal quantity of labor. In fact, this conclusion is simply incorrect. A glance at Figure 3 readily confirms that the firm in the model will employ an inefficiently small quantity of labor for whatever level of output it chooses to produce. The reason, of course, is that the condition $R_1 = r_1$ is

---

13 A more "realistic" example is provided by the revenue function $R(x_1, x_2) = 2ax_1^b + 4bx_2^c + 2cx_1$, $a > 0$, $b > 0$, $c > 0$. The argument is similar to that just given.

14 After the paper was completed it came to our attention that Westfield had, in fact, presented results essentially the same as those derived in this section (See [12], p. 430). His verbal statements involved some slight inexactness, and there appears to be an error—probably typographical—in the footnote where he demonstrates the effect of regulation on the profit-maximizing output level. However, there is no doubt that he had arrived correctly at the relevant conclusions.
consistent with the efficiency requirement \( \frac{R_1}{R_2} = \frac{r_1}{r_2} \) if and only if \( R_1 = r_1 \), which will occur only at the unconstrained profit-maximizing point.

Consequently, one may wish to inquire further into the effect of the regulatory constraint upon the firm’s use of labor. Does the firm’s employment of labor necessarily increase (or necessarily decrease) when a regulatory revenue constraint is imposed? As is the case in its use of capital, is there necessarily a monotonic relationship between the quantity of labor hired and the gap between \( s \) and \( r_1 \)? It can be shown that no such generalization is possible. Depending on the marginal rate of substitution between \( x_1 \) and \( x_2 \), the employment of labor may either increase or decrease as \( s \) gets closer to \( r_1 \). If labor and capital are complementary in the gross revenue function, then as the quantity of capital used by the firm increases its use of labor will also rise. But if capital is a substitute for labor in producing revenue, then \( x_1 \) and \( x_2 \) will move in opposite directions in the A-J model.

The proof follows the comparative-statics procedure utilized in deriving the Takayama result. However, the proof will be indicated only implicitly as part of our discussion of another issue—the effects of the regulatory constraint on the firm’s output level. It will be recalled that the A-J allegation on this subject, designated proposition 4 above, asserts that effective rate-of-return regulation will lead the firm to increase its output level. Let us now see under what circumstances this is true.

Using the production function \( z = z(x_1, x_2) \), one obviously has

\[
\frac{dz}{ds} = \frac{dx_1}{ds} + \frac{dx_2}{ds}
\]

where \( z_1 \) and \( z_2 \) are, respectively, the marginal products of capital and labor. To determine the sign of \( \frac{dz}{ds} \) it is necessary to evaluate \( \frac{dx_1}{ds} \) and \( \frac{dx_2}{ds} \). Total differentiation of (13) yields

\[
R_{x_1}dx_1 + R_{x_2}dx_2 = 0 \quad \text{or} \quad dx_2 = -\left(\frac{R_{x_1}}{R_{x_2}}\right)dx_1.
\]

Since \( R_{x_2} < 0 \) by the strict concavity of \( R \) or simply by diminishing (marginal revenue) returns to labor, it follows immediately that the change in the employment of labor, \( dx_2 \), must have the same sign as the change in the use of capital, \( dx_1 \), if and only if capital and labor are complements in the production of revenue, \( R_{x_1} > 0 \). If they are substitutes in revenue production, their employment levels must change in opposite directions.

Substituting from (18) for \( dx_2 \) into (17), one obtains

\[
\frac{dz}{ds} = \left(\frac{z_1 - z_2}{R_{x_2}}\right)\frac{dx_1}{ds} + \frac{z_1 R_{x_2} - z_2 R_{x_1}}{R_{x_2}} \frac{dx_1}{ds}.
\]

We want to evaluate \( z_1 R_{x_2} - z_2 R_{x_1} \) in (19). For this purpose we note that since \( R = \) total revenue = \( pz \),

\[
R_x = (p + p'z)z_1 \\
R_{x_2} = (p + p'z)z_{x_2} + (2p' + p''z)z_2 \\
R_{x_1} = (p + p'z)z_{x_1} + (2p' + p''z)z_1 z_2.
\]
Consequently,
\[ z_1 R_{21} - z_2 R_{21} = (p + p'z)z_1z_{22} + (2p' + p''z)z_1^2z_1 
- (p + p'z)z_2 z_{21} - (2p' + p''z)z_2 z_1^2 z_1 \]
or
\[ z_1 R_{22} - z_2 R_{21} = (p + p'z)(z_1 z_{22} - z_2 z_{21}). \] (20)

Now, \( p + p'z \) is simply \( MR_c \), the marginal revenue of \( z \), and is usually assumed to be positive (though with a constraint on rate of return this, clearly, need not be so). We may also assume \( R_{22} > 0 \) (diminishing marginal revenue returns to labor—which follows from the strict concavity of \( R \)). As a result, by (20) and (19) \( dz/ds \) must have the same sign as \( dx_1/ds \) (i.e., they must both be negative) if and only if
\[ z_1 z_{22} - z_2 z_{21} < 0. \] (21)

There is, however, no basis on which to prejudge the sign of \( z_{21} \) since capital and labor can be either complements or substitutes in producing output. Therefore, we simply cannot conclude that the A-J proposition 4 is always valid.

Yet there is a plausibility argument in its favor. Since we do expect \( z_1 > 0 \) and \( z_2 > 0 \) (positive marginal productivity of each input) we can write (21) as

\[ \frac{z_{22}}{z_2} < \frac{z_{21}}{z_1}. \]

But this requires that the proportional change in the marginal physical product of labor which results from an increase in labor alone is algebraically smaller than the proportional change in the marginal physical product of capital which results from the same increase in labor alone. The condition is met if \( z_{21} > 0 \) so that capital and labor are complements in production. If, on the other hand, \( z_{21} < 0 \) so that capital and labor are substitutes, then it is met if the proportional decrease in marginal physical products as a result of a labor increase is greater for labor than for capital. In sum, the expression \( z_1 z_{22} - z_2 z_{21} \) can be said to be negative so long as labor is a better substitute for labor than it is for capital.\(^{17}\)

Alternatively, by differentiating totally the equilibrium conditions for an unconstrained profit-maximizer, that is, (3a) and (5b) with \( \lambda = 0 \), and the production function for the firm, all with respect to a change in the market cost of capital, \( n \), one can easily show that the key expression in (19), \( z_1 R_{22} - z_2 R_{21} \), will be positive if and only if \( x_1 \) (capital) is a “regressive” input—i.e., an input such that a rise in its price leads the firm, ceteris paribus, to increase the output of the final product. In short, if marginal revenue is positive and if labor is a better substitute for labor than it is for capital—or equivalently, if capital is not a regressive factor—then \( dz/ds < 0 \) so that output will indeed increase as the regulatory constraint becomes tighter (s decreases) as is required for the validity of the A-J conjecture (proposition 4 above).

---

\(^{17}\) If both terms in (21) are multiplied by \( x_0 \), the expression takes the form of a comparison of elasticities of marginal products with respect to changes in \( x_0 \) and it is then units free.
Finally, we are now in a position to deal with (alleged) proposition 2, which claims that the profit-maximizing firm subject to an effective rate-of-return constraint will necessarily employ a capital-labor ratio higher than that of an unconstrained enterprise.

We will first show in general terms why this need not be true and then we will offer a specific counterexample.

Proposition 2 asserts that

\[
\frac{d(x_1/x_2)}{ds} < 0 ,
\]

that is, by (16)

\[
\text{sign} \left( \frac{d(x_1/x_2)}{ds} \right) = \text{sign} \left( \frac{dx_1}{ds} \right).
\]

(22)

Since we care only about signs we can look at the proportional change in \( x_1/x_2 \) when \( s \) decreases (i.e., when regulation is tightened).

We have as this proportional change

\[
\frac{d(x_1/x_2)}{x_1/x_2} = \left( \frac{dx_1}{x_1} - \frac{x_1}{x_2} \frac{dx_2}{x_2} \right) = \frac{dx_1}{x_1} - \frac{dx_2}{x_2}.
\]

But from (18) we know that

\[
dx_2 = -(R_{12}/R_{22})dx_1.
\]

Therefore

\[
\frac{d(x_1/x_2)}{x_1/x_2} = \frac{dx_1}{x_1} + \frac{R_{12}}{R_{22}} \frac{dx_1}{x_2} = dx_1 \left( \frac{1}{x_1} + \frac{1}{R_{12}} \right).
\]

Thus (22) holds if and only if

\[
\left( \frac{1}{x_1} + \frac{1}{R_{12}} \right) > 0 , \quad \text{i.e.,} \quad \frac{x_2}{x_1} > -\frac{R_{12}}{R_{22}}.
\]

Now this is surely true if \( R_{12} < 0 \), i.e., if capital and labor are substitutes in the production of revenue, because then we have \( R_{22} < 0 \), \( R_{12} < 0 \).

But suppose \( R_{12} > 0 \). Then the preceding condition becomes

\[
\frac{x_2}{x_1} > \frac{R_{12}}{R_{22}} \quad \text{or} \quad \frac{|R_{12}|}{|R_{22}|} > \frac{x_2}{x_1} > \frac{R_{12}}{R_{22}}
\]

which is obviously not necessarily true, nor is it even easily interpretable.

We now provide a simple counterexample to the allegation of proposition 2. Let us suppose the firm’s revenue function is

\[
R(x_1, x_2) = ax_1^2/2 + bx_1x_2 + cx_2^2/2
\]

where

\[
a < 0 , \quad c < 0 \quad \text{and} \quad ac - b^2 > 0
\]

so that \( R(x_1, x_2) \) is strictly concave. By (5b) our constrained profit maximizer will satisfy \( R_2 = r_2 \), i.e.,

\[
bx_1 + cx_2 = r_2 \quad \text{or} \quad x_2/x_1 = r_2/cx_1 - b/c
\]
which increases as $x_1$ increases since $r_1 > 0$, $c < 0$. But by proposition 5, $x_1$ will be larger for the constrained than for the unconstrained firm, so that in this case the labor-capital ratio will be increased by the regulatory constraint, contrary to the assertion of proposition 2.

7. The revenue-maximizing firm under regulation

We have now completed our discussion of the basic A-J propositions and their formal validity. We turn now to some related theoretical results that have developed from the model. One line of investigation which has grown out of the A-J analysis has examined the effects of the regulatory constraint (2) upon the firm that seeks to maximize total revenue or some other maximand instead of total profit. Such models have been studied by Bailey and Malone [3], Kafoglis [5], and Zajac [14]. Among the maximands that have been suggested by these authors are total revenue (sales), $p$, physical volume, $x$, and rate of return, $\pi/x$. In each case the analysis proceeds exactly as before, only with the substitution of one of the preceding maximands for the total profit function (1). There is no need to go through the details of the analyses.

Not surprisingly, Zajac concludes that for a regulated firm maximizing its rate of return, the capital-labor ratio becomes indeterminate. Any point on the constraint curve in Figure 3 becomes as good as any other. The firm will simply select any capital-labor combination that enables it to attain that iso-rate-of-return curve, since it will then have achieved the greatest rate of return permitted by regulation.

Bailey and Malone, however, following the A-J procedures find that under sales (i.e., total revenue) maximization the nature of the inefficiency discussed by Averch and Johnson appears to reverse itself. Assuming that the firm’s production function is quasiconcave so that the iso-product curves are convex to the origin, they prove:

**Proposition 6.** The sales-maximizing firm under rate-of-return regulation is motivated to use a labor-capital ratio greater than that which minimizes cost for the output level it chooses to produce.

To prove proposition 6 we note that to obtain the new maximand, total revenue, we simply drop the terms $-r_1x_1$ and $-r_2x_2$ from the profit function (1) so that (5a) and (5b) become respectively

\[ (1 - \lambda) MR_1, x_1 = -\lambda r_1 - \lambda (s - r_1) \]
\[ (1 - \lambda) MR_2, x_2 = -\lambda r_2 . \]

Dividing the first equation by the second we have

\[ z_1/z_2 = r_1/r_2 + (s - r_1)/r_2 \quad (23) \]

where the last term is obviously positive. Comparison with (7) shows that, unlike the profit-maximizing case, this time $z_1/z_2$ exceeds $r_1/r_2$ which, given the convexity assumption for the iso-product curve, can only occur at a point where the labor-capital ratio exceeds its efficient value.

---

We had intended to include in this paper a discussion of Klevorick’s graduated tax return proposal [7]. However, C. R. Wachter [13] has found a significant error in the original formulation and it has therefore been thought advisable to postpone the discussion until the matter has been resolved.
This is shown explicitly in Figure 3. Suppose the firm's output corresponds to iso-product locus $QQ'$. At every point on that locus, since output is fixed, the price of the company's product is the same. Hence at any point on $QQ'$ the value of $z_1/z_2$ is given by the slope of this curve. At the efficient point, $K$, this slope is equal to $r_1/r_2$, the slope of the budget line. Only at points on $QQ'$ above and to the left of efficient point $K$ will $z_1/z_2$, the slope of $QQ'$, exceed $r_1/r_2$. But at these higher points on $QQ'$ the labor-capital ratio exceeds its efficient value.

To compare geometrically the sales-maximizing firm's behavior with that of the profit-maximizing firm, assume that demand is elastic over the relevant range. In this case the firm will move to the highest iso-product (iso-revenue) locus attainable on the constraint locus.

Usually in sales (total revenue) maximization models the firm is taken to maximize its total revenue function $R(x_1, x_2)$ subject to some sort of profit constraint such as $\pi(x_1, x_2) \geq mx_1$, which guarantees that the firm takes advantage of opportunities to earn enough profit to satisfy its stockholders. The discussion implicitly assumes that the profit constraint (2) of the regulatory model also serves, with the inequality reversed, as the minimum profit requirement for the sales-maximization hypothesis. If the regulatory constraint permitted the firm to earn more than the profit needed to satisfy stockholders, revenue maximization would require it to move to an iso-rate-of-return curve lying entirely outside the shaded region in Figure 3, where the shaded region represents the input combinations prohibited by the fair rate of return ceiling.

Since the firm wishes, therefore, to move to the highest iso-revenue locus on the constraint curve, it will go to the locus tangent to the constraint curve. And since any iso-revenue (iso-product) curve is negatively sloped, the input point chosen by the regulated sales maximizer will therefore occur on a negatively sloping portion of the constraint curve. It must then involve a labor-capital ratio higher than that selected by the profit-maximizing regulated firm which, as we have seen, corresponds to the rightmost (vertical) point on the constraint curve (point $\pi_{\text{reg}}$ in Figure 3).

The preceding result, since its conclusion is precisely the opposite of that of proposition 3, may be considered one of a number of attempts to show that the A-J conclusions lack general validity.

A rather informal argument by Zajac with a somewhat similar objective may be mentioned in passing. This discussion raises no questions about the formal validity of the deductions from the theoretical model but does indicate doubts about the model's applicability. Zajac suggests that the firm in reality knows its cost relationships far better than its demand function. As a result it may find it far easier to locate points on its efficient locus than to approximate its constrained profit maximum point, $\pi_{\text{reg}}$ in Figure 3. Zajac suggests that the firm may perhaps attempt to work towards $\pi_{\text{reg}}$ by any of a large number of alternate trial-and-error procedures. He chooses to focus on two of these possible series of moves. In the first set the firm follows along the regulatory constraint curve, while in the other

---

19 Since we take $p = p(z)$, for any fixed value of $z$ we must also have a constant value of $p$. Consequently, a locus along which $z$ is fixed (an iso-product locus) must also involve a fixed total revenue ($p_2$).
the firm, knowing no better, proceeds to the point of minimum cost for whatever output it happens to be producing initially (a point on the efficient path) and then moves roughly along the efficient path toward the most profitable point on that path, the point where it intersects the constraint curve. Which of these strategies, if either, the firm is in fact likely to adopt we cannot say. Zajac points out, though, that depending on the relative speed of the two trial-and-error processes, it is perfectly possible that the efficient-path strategy will actually prove more profitable for the firm, despite its relative unprofitability in static equilibrium with full information. Clearly, if this is the course selected by management, the conclusion of proposition 3 simply does not hold. Of course, the Zajac argument shows only that with imperfection of information of the sort that may plausibly be expected in this area it is virtually impossible to predict what will happen. Things may turn out either better or worse than the A-J model suggests. But that is just the point that Zajac wishes to make.

8. Regulatory lag

In practice the rate-of-return constraint is not enforced continually. Rather, the regulator restudies the company's prices and profits from time to time and declares, if necessary, that these be readjusted so as to yield for the firm the fair rate of return. Regulatory lag means, for example, that the firm will often be able to earn profits greater than those considered appropriate by the regulator, as judged by the permitted rate of return, for limited periods of time. We will argue later that this lag is potentially of very great significance for economic efficiency in general. In the present context one might surmise that it serves to weaken the A-J input-proportion effect, at least to some extent.

Bailey and Coleman [2] have undertaken to investigate formally some of the implications of regulatory lag for the A-J model. They discuss two possible courses of events, one in which the firm simply acquires inefficiently large quantities of capital and holds them idle until the regulator, by changing the price the firm can charge, readjusts the total return the firm is permitted to earn, and the second in which the company makes the best possible use of that capital even though the quantity is inefficiently large for the output being produced. Since the former case, though the less realistic, is considerably easier to describe, and since it enables us to characterize the essence of the Bailey-Coleman argument, we will use it as the basis of our discussion. Moreover, although Bailey and Coleman considered the case of a firm with a finite horizon as well as the case of a firm with an infinite horizon, we shall restrict our discussion to the latter as it serves to bring out the basic point of their argument.

The course of events envisioned in the Bailey-Coleman article is roughly the following: capital can be acquired easily, but the regulator does not adjust the company's prices immediately after the acquisition of new capital. The firm must therefore consider the advisability of obtaining an inefficiently large quantity of capital à la A-J, knowing that the new capital will reduce profits until the price adjustment and increase profits thereafter. While management does not know the length of the time it takes for the regulator
to react by making a price adjustment, it does have an estimate of its expected value, believing this to be \( n \) periods in the future.

Utilizing the same notation as before, if the excess capital is acquired at once, the company will immediately incur a cost of \( r_1 \Delta x_1 \) dollars per period, where \( \Delta x_1 \) is the quantity of capital in question. Once the regulatory readjustment takes place the company will add to its total earnings \( s \Delta x_1 \) dollars per period but will continue to have to pay \( r_1 \Delta x_1 \) dollars per period for the privilege.

Letting \( D \) be the appropriate discount factor, the net effect upon the profits of the firm will therefore be

\[
\Delta \Pi = - r_1 \Delta x_1 \sum_{t=0}^{n-1} D^t + (s - r_1) \Delta x_1 \sum_{t=n}^{\infty} D^t.
\]

This will be positive, (i.e., the investment will be profitable) if and only if

\[
- r_1 \sum_{t=0}^{n-1} D^t + (s - r_1) \sum_{t=n}^{\infty} D^t > 0, \quad \text{i.e.,}
\]

\[
- r_1(1 - D^n)/(1 - D) + (s - r_1)D^n/(1 - D) > 0,
\]

or

\[
- r_1 + sD^n > 0.
\]

Since \( D < 1 \), the last inequality shows that if the length of the lag is sufficiently great relative to \( s - r_1 \) it will notpay to overinvest by the postulated amount \( \Delta x_1 \). Increased regulatory lag will therefore decrease the likelihood in this model that the firm will want to acquire an excessive quantity of capital in the manner suggested by A & J.

In this simple and unrealistic case, the one in which the excess capital is held idle, the amount of excess capital is not well defined because there are no diminishing returns to excess capital. The limit to the amount that will be acquired is presumably imposed by an influence that does not appear in the model explicitly. If the elasticity of demand for its product limits the amount that the company can, in fact, earn to less than the amount permitted by an expanded rate base, there simply will be no point in acquiring additional capital.

The value of \( \Delta x_1 \) that enables the firm to reach the point of maximum profits under the fair rate of return constraint \((\pi_{\text{reg}} \text{ in Figure 3})\) will then be the amount, \( \Delta x_1^{*} \), that it pays the company to obtain. If \( n \) is sufficiently small to make it profitable to acquire any excess capital at all, it will pay to obtain this full amount, \( \Delta x_1^{*} \).

Bailey and Coleman show, however, that this all-or-none result is an artifact of the simplifying premise that the excess capital is retained unused. They argue that it will, in fact, always pay to use the capital to replace some of the company’s labor, rather than keeping it idle. In that case we may expect a diminishing profit yield to increased quantities of capital as a consequence of the diminishing marginal rate of substitution of capital for labor. Bailey and Coleman show that in their model it will always pay the firm to acquire some excess capital, but that the amount it is profitable to obtain will vary inversely with the length of the regulatory lag. The reason is simple. In their world the excess capital always faces a period in which it is expected to incur a net loss. The sooner it can begin to earn a net profit as a result of regulatory price readjustment, the more capital it will pay the firm to acquire for the purpose.
The authors of this paper have also constructed and examined a number of models of regulatory lag. The regulatory process and its effects as depicted in our model are, however, somewhat more complex than in most of the other models that have been described, in accord with our view of the nature of the regulatory process in reality. In our model the time sequence of profits and losses is reversed from that in the Bailey-Coleman model. While Bailey and Coleman regard the period before a regulatory review as the time when the firm suffers a loss because it is carrying an excessive amount of capital, now the period between reviews is regarded as the time when the firm has the possibility of earning a profit rate exceeding that specified by the constraint. When the regulatory review occurs, this excess is eliminated by the regulators' adjustment of the prices the firm can charge. In contrast, in the Bailey-Coleman model it is the review point which enables the firm to gain at last the benefits of the excess capital it has purchased earlier.

As we see it, regulation proceeds, roughly, by taking the latest information available on company costs and demands, and resetting prices so that the firm's current rate of return—based on its current capital stock, technology, demand conditions, and so forth—is reduced to the "fair" rate level. The firm is required to leave those prices unchanged until another regulatory review takes place either at the behest of the company or some other interested party. Thus, unlike the company in the A-J model, the firm in this analysis cannot simply select an output level and charge for it the corresponding price as indicated by the demand curve. On the contrary, the price is set by the regulators and the company, as a public utility, is then required to supply whatever volume of output happens to be demanded at that price.

How then can any excess profits be earned? Leaving aside any demand shifts the firm may be able to induce with the aid of advertising and the like, the answer is that by research and innovation the company can hope to reduce its costs below those that prevailed when prices were last set by the regulators. So long as the initial output and input prices prevail, it will be able to retain for its stockholders whatever savings the innovation makes possible. But this additional earnings flow is only a temporary benefit. At the next regulatory review, prices will be readjusted once again to take into account the firm's improved technology and the process begins all over again.

In the analysis of this process the firm is taken to utilize (at least) three inputs: labor, capital, and knowledge. Labor is a non-durable input, and the amount employed is determined anew at each of the firm's decision points. Each instant's net investment, however, constitutes an addition to the firm's total stock of capital, and similarly, each period's research adds to the firm's accumulated stock of knowledge. The issues, then, are: first, whether or not in this model the regulated firm will be motivated to employ an efficient combination of the three inputs; second, whether the degree of input inefficiency, if there is any, varies with the length of the regulatory period; and third, how the equilibrium stock of knowledge is affected by the length of time between regulatory reviews.

This last issue is potentially of great importance. If one can demonstrate a functional relationship between the length of time between regulatory reviews and the long-run equilibrium stock of
knowledge, it would suggest that regulators have at their disposal an instrument, regulatory lag, of which they have not made much conscious use. Suppose, for example, that the regulated firm faces a static demand curve—i.e., one which does not shift with time. Now, the long-run goal of regulation is surely related to the net flow of benefits to consumers and to the producer (the total producer’s plus consumers’ surplus produced by the regulated firm) and the distribution of those benefits between producer and consumers. Other things being equal, the total net benefits generated by the regulated firm will be increased by downward shifts in the firm’s long-run marginal cost curve produced by improvements in its technology—i.e., by increases in the long-run equilibrium stock of knowledge.

Suppose the equilibrium stock of knowledge increases with the length of the regulatory lag, because the firm undertakes more research the longer are the periods during which it enjoys the temporary benefit of an excess earnings flow. Then the total discounted present value of the benefits generated by the regulated firm will rise when the length of the regulatory lag increases. For example, with instantaneous regulation—a lag of length zero—and technical progress which increased the productivity of capital and labor proportionately, one would expect the firm to undertake no research outlays since it could never reap any benefits from such an expenditure, while with no regulatory review—a lag of infinite length—one would expect the firm to accumulate a maximal stock of knowledge since it reaps the benefits of its research work forever.

While one might consequently expect the length of the lag to be positively related to the total discounted present value of producer’s plus consumers’ benefits, it is clear that a smaller proportion of these benefits will be distributed to the firm the shorter is the lag. That is, the shorter the periods during which the firm can earn “excess” profits and retain them for itself, the greater is the portion of the accumulated surplus which is taken from the firm and given to consumers.

To each length of regulatory lag, then, there corresponds a discounted present value of total gains and a proportion of these gains which accrues to the consuming public. The latter decreases as the lag increases, while one would expect the former to increase as the lag increases. In the simplest case, society faces an opportunity locus of the type shown in Figure 6, where total surplus (benefit) is measured on the vertical axis and the proportion of the total gains accruing to consumers is measured on the horizontal axis. Point A corresponds to instantaneous regulatory adjustment, while point B corresponds to an infinitely long period between rate reviews. Society—or the regulators—may then be assumed to choose among the points on this locus with the aid of a social welfare function defined on the two arguments—total net benefits, and their distribution—and to maximize the social welfare function subject to the opportunities-locus constraint. In the simplest case, the optimal point selected in this way will correspond to a unique length of the regulatory lag. In any event, the regulators can then choose a length for the period between reviews which will attain the social optimum with respect to the amount and distribution of producer’s plus consumers’ surplus. If regulatory agencies have chosen the time period between
reviews in some more or less *ad hoc* fashion, this model would suggest that they have neglected a potentially valuable instrument of control.

With the aid of some additional notation, it is easy to describe one of the formal models used to investigate this issue of regulatory lag and research. Let

\[
\begin{align*}
  z &= \text{the flow of output} \\
  x_1 &= \text{the stock of capital} \\
  x_2 &= \text{the quantity of labor used} \\
  x_3 &= \text{the stock of knowledge} \\
  x_4 &= \text{the flow of investment} \\
  x_5 &= \text{the flow of research activity},
\end{align*}
\]

all at time \( t \),

\[
p_i = \text{the price the firm is allowed to charge in the } i\text{th regulatory period; } p_i \text{ is exogenously given while for } i \geq 2, p_i \text{ emerges as the adjusted price after the } (i - 1)\text{th regulatory review,}
\]

\[
z = F(x_1, x_2, x_3) \text{ is the production function,}
\]

\[
r_2 \text{ is the wage rate,}
\]

\[
g(x_1) \text{ is the (total) cost of investment,}
\]

\[
h(x_3) \text{ is the (total) cost of research and,}
\]

\[
v \text{ is the discount rate.}
\]

Assume that the length of time between regulatory reviews, i.e., the length of the regulatory lag, is \( T \).

The discounted present value of the firm’s profits is then

\[
\Pi = \sum_{i=1}^{\infty} \int_{(i-1)T}^{iT} [p_i F(x_1, x_2, x_3) - r_2 x_2 - g(x_1) - h(x_3)] e^{-vt} dt. \quad (24)
\]

The firm’s goal is to maximize the discounted present value of profits subject to the transition equations (25) and (26) for capital and the stock of knowledge:

\[
\frac{dx_1}{dt} = x_1 \quad (25)
\]

\[
\frac{dx_2}{dt} = \frac{dx_3}{dt} = \frac{dx_4}{dt} = \frac{dx_5}{dt} = 0
\]

\[
\frac{dx_1}{dt} = \frac{dx_2}{dt} = \frac{dx_3}{dt} = \frac{dx_4}{dt} = \frac{dx_5}{dt} = 0
\]
\[
\frac{dx_3}{dt} = \dot{x}_3 \geq 0. \tag{26}
\]

As is clear from (25) and (26), we assume that there is no capital depreciation and no decay of knowledge through time. Finally, the firm must operate subject to the transition equation for price which is dictated by the price-adjustment procedure of the regulators, namely,

\[
p_n(x_2)F(x_1[(i - 1)T], x_3, x_3[(i - 1)T]) - r_2x_2 = sx_1[(i - 1)T]. \tag{27}
\]

This price-transition equation represents the fair-rate-of-return constraint in our model incorporating regulatory lag. It states that the regulators proceed as follows in setting the price for period \(i\). They assume that the firm does no research and no investing after the end of period \(i - 1\), so that its capital stock and stock of knowledge in period \(i\) are assumed to be \(x_1[(i - 1)T]\) and \(x_3[(i - 1)T]\), respectively. Then, the regulators calculate what price will just yield to the firm the fair rate of return in period \(i\) on its assumed capital stock for that period. In fact, equation (27) may be regarded as an implicit equation for \(x_2\) since \(x_1[(i - 1)T], x_3[(i - 1)T], s\) and \(r_2\) are all given at the beginning of period \(i\). When the solution value for \(x_2\) is obtained, \(p_n\), the price the firm is allowed to charge in period \(i\), follows from the inverse demand function, \(p = p(F(x_1, x_3, x_3))\).

We are still investigating the implications of this model, and others similar to it, for the issues described earlier. The methods of analysis used to examine these questions are, primarily, the standard tools the economist employs in studying problems involving optimization over time: optimal control theory, dynamic programming, and nonlinear programming. In our case, we use these techniques to characterize the optimal path for the firm maximizing the discounted present value of profit subject to the particular formulation of the regulatory constraint. For example, in the model just presented, we use these methods to study the optimal path for the firm maximizing (24) subject to (25)–(27). Then one examines the conditions characterizing this path to see whether or not the firm’s use of inputs along its optimal trajectory is efficient. In particular, the efficiency or inefficiency characterizing the equilibrium point to which the firm’s optimal path leads it is also examined. To investigate the effect of the length of the regulatory lag on the firm’s behavior, an experiment in comparative dynamics is then performed. The parameter representing the regulators’ speed of response—in the model just presented, this parameter is \(T\)—is varied, and one observes the effect of this variation on the path chosen by the firm. In particular, one performs a special comparative-statics experiment at the firm’s equilibrium point to determine the effect of the change in the regulatory lag on the equilibrium stock of knowledge.

The main reason for presenting the model here is that it appears, despite its oversimplification, to be a closer representation of the regulatory process than has been available before and because it encompasses the important policy issues that were just discussed. While results are still being derived for this and similar models, it would be fair to say that the general impression suggested by the conclusions already obtained is that there is a definite role for regulatory lag to play in inducing regulated firms to undertake research.
but that the exact relationship—even the direction of the relationship—between the length of the lag and the equilibrium stock of knowledge depends critically on the nature of the production function $F(x_1, x_2, x_3)$ and, particularly, on the way the stock of knowledge enters that function. A similar remark applies to the effects of regulatory lag on the A-J conclusions. In some models, notably the Bailey-Coleman model, the inefficiency in the capital-labor ratio declines as the lag increases. But here too there is a great deal yet to be done.

9. Evaluative comments

As a contribution to the theory of regulation the work of Averch and Johnson represents a significant landmark. It shows how rigorous methods can be used to draw illuminating conclusions in an area in which historical description and impressionistic discussion has until recently played a preponderant role in the literature. This is clearly to the good, and it is hard to see how this accomplishment can be questioned.

This still leaves open, however, the issue of the significance of the A-J results for policymaking. And here, it seems to us, the conclusion we should draw is rather mixed. The analysis is certainly pertinent in pointing out that rate-of-return regulation can impose social costs in the form of input inefficiency. Unfortunately, no one has yet found a good substitute for limitation on rate of return as a device to prevent the earning of monopoly profits by public utility firms, an objective that few persons probably would be willing to abandon. Yet any constraint on the decision process which is not designed explicitly to avoid undesired interference with the allocation of resources is likely to produce distortions in their usage. Most taxes have such effects, and it is hardly surprising that a fair-rate-of-return constraint will have such consequences.

The next step, then, is to determine the most important of these undesirable effects and to consider what can be done about them. The A-J effect certainly seems to be a candidate, although we have seen that in any particular model it may or may not be present, depending on the behavioral assumptions employed. Thus, for example, in the case where the firm is a sales maximizer we have seen that excessive labor rather than capital intensiveness follows from the premises. But even in a world populated by profit-maximizers we cannot be sure that the consequences for public welfare of the A-J effect are likely to be very serious. Harberger, in a well-known article [4] has estimated, on the basis of a calculation that was admittedly crude, that the social loss through misallocation resulting from monopoly and elements related to it is surprisingly small—a rather negligible proportion of the GNP. Leibenstein [9] following this line of thought has suggested that a far more serious source of welfare loss is what he calls "X inefficiency"—cases where management, characteristically as a result of imperfect knowledge, has simply failed to organize the firm's operations in a highly efficient manner.

Regulation seems to provide particularly apt examples of this distinction. The A-J overcapitalization is an example of the inefficiencies emphasized in the more conventional analyses. But even if it occurs in practice it does not seem likely to produce effects that are very serious. This seems so, if for no other reason, because, as we
have already observed, firms may have neither the extensive information nor the refined decision processes necessary to lead unerringly to the A-J input distortions.

Of course, even if the A-J overcapitalization effect is not widespread, it does not follow that rate-of-return regulation is without its dangers. It has been suggested that it has sometimes been used by regulators as a means to keep inefficient firms in business, shoring them up by preventing their rivals from passing their lower costs on to the consumers. The regulator's apparent distaste for the demise of any business firm no matter how tenuous its economic justification may well contribute in no small measure to X inefficiency.

Moreover, a rate-of-return ceiling by its very nature tends to contribute another X-inefficient influence. By ruling out all profits in excess of some predetermined rate as potential fruits of monopoly, a rate-of-return constraint also precludes extraordinary rewards for extraordinary entrepreneurial accomplishment. Indeed, if such a constraint were applied with complete inflexibility it might well eliminate altogether any financial reward for efficiency and innovation. If the firm cannot hope to earn more than the ceiling, even temporarily, and if the demands for its products are inelastic so that it can reasonably count on that profit rate in any event, what can management hope to gain by working for increased efficiency in its operations? Fortunately, as we have indicated, regulatory lag, like the Schumpeterian innovation process, by permitting temporary rewards for efficiency can perhaps provide just the necessary incentives. This is not the place to pursue the issue further. The point is simply that while regulation may well be suspected of being the source of some non-negligible inefficiencies in the economy, it is not clear that the phenomenon encompassed by A-J analysis is the most disquieting of these.

References


---

26 On this see see also the alternative theoretical analyses of some other problems of rate-of-return regulation in Wells and in Westfield.


