1977

The Economics of Bureaucracy

Ronald Wintrobe

20188

RESEARCH REPORT 7714

THE ECONOMICS OF BUREAUCRACY

by

Ronald Wintrobe

November 1977

The Economics of Bureaucracy*

Ronald Wintrobe

University of Western Ontario

* This paper is based on my 1976 doctoral dissertation at the University of Toronto. Earlier versions were presented in seminars at the University of Chicago and the University of Western Ontario. I am indebted to Albert Breton, Yehuda Kotowitz, and Allan Hynes for innumerable discussions, and to George Stigler for constructive comments on a previous draft. Any errors remain my responsibility.

I. Introduction

Managers in organizations are usually thought of as having "control" over the decisions and actions of their subordinates, or, alternatively, as having "power" or "authority" over them. This control is commonly thought to be exercised by giving orders or instructions to subordinates, or by the use of rules, which are simply generalized instructions, in that they apply to a number of subordinates simultaneously, and they persist over time unless explicitly revoked or changed. On the other hand, in almost all organizations, rules are not always followed, orders are not always obeyed, and subordinates may often be observed to act without orders, i.e. on their own authority. Moreover, the extent to which subordinates exercise their own "discretion" appears to vary systematically, both between employees at different hierarchical levels of the same organization, and among employees at the same level in different organizations with different characteristics.

This paper develops an economic model to explain and to predict the equilibrium level of employee discretion or its obverse, the degree of "routinization" or working time, defined as the extent to which subordinates act according to instructions or rules set by their superiors rather than on their own authority.

Webster's New Collegiate Dictionary defines "bureaucracy" as "the use of routine procedure in administration", and one of our purposes is to show why it is in the nature of hierarchies to routinize the working time of their employees.[1/] The major purpose of the paper, however, is to focus on differences in discretion among organizations in order to illuminate and predict differences in organizational incentive systems, structure

(size, span of control, etc.) and behavior. Perhaps the most striking feature of the model is its use of simple economic theory to illuminate many aspects of bureaucratic behavior. For example, the analysis suggests a simple economic interpretation of authority and obedience, and one, moreover, which has considerable predictive power. Other, less basic, but more puzzling aspects of bureaucratic behavior, such as the tendency of bureaus to ossify, and to continually re-organize themselves, the fact that, in many organizations, subordinates are never dismissed even in cases of obvious incompetence, red tape, organizational "slack", and other phenomena are shown to be easily understood with the aid of simple tools of economic theory.

The plan of the paper is as follows: section 2 shows how incentives may be used to routinize the working time of employees in the context of the simplest and most authoritarian organizational incentive system., one in which subordinates exercise no discretion. The characteristic inefficiencies of routinization are pointed out, and it is thereby shown why organization managers typically provide incentives to subordinates to exercise their own discretionary authority. Section 3 shows how these incentives work.

In section 4 we focus on the organization manager, and his choice of the optimal incentive system. The model developed there predicts the equilibrium use of different bureaucratic incentives-- and therefore, the equilibrium level of discretion as a function of basic technical organizational characteristics such as technology, the costs of supervising employees, the elasticity of substitution between factors, and other variables. In particular, we derive an inverse relation between the average level of discretion and organizational size and show that this relationship provides a limit to the size of the firm.

2. Rules as prices: the "pure authority" system

An economic theory of bureaucratic behavior i.e, of the behavior of subordinates in a hierarchy, explains differences in behavior by differences in opportunities rather than in preferences. If bureaucrats do not differ systematically in their preferences from other agents in economic theory, they will follow rules when it is in their individual interest to do so, and not otherwise.

Consequently, to achieve the compliance of subordinates to rules and hierarchical directions, organizations must devote resources to monitoring subordinates' actions, and there must be sanctions for non-compliance. The instruments required for this purpose are identical to those used by the state in Becker's (1968) economic model of law enforcement. To show how they operate, let p be an index of how closely a subordinate's activities are monitored (i.e. the probability that he will be "caught" violating instructions); f the sanction imposed for non-compliance; g an index of the personal gain (monetary or non-monetary) to the subordinate from violating rules or instructions, and Y_0 the base wage necessary to attract the appropriate quality of personnel.

The expected utility theorem implies that subordinates will obey rules and instructions as long as the utility of the certain income from doing so is greater than the expected utility from disobeying them, i.e. as long as

$$(1) \qquad U(Y_0) > pU(Y_0 - f) + (1 - p)\, U(Y_0 + g)$$

Using this framework, the decision of a subordinate to follow organizational rules or direct instructions amounts to his acceptance of an insurance policy. Equation (1) states simply that for the insurance to be consistently accepted, it must be offered on favorable terms. This may be

achieved by raising p and f to sufficiently high levels that the expected gain to subordinates from non-compliance is always negative.

Now, in all but the most routine bureaucracies, "exceptional" situations will arise from time to time, i.e. situations which do not appear to a subordinate to be covered by existing rules, where the rules appear to be in conflict, or where it may seem obvious that the rules are clearly inappropriate. In any of these circumstances, a subordinate is faced with the choice of acting on his own or consulting his superior. Again, the subordinate's decision may be analyzed as a choice between "buying" insurance or taking a gamble, as in equation (1). Moreover, any superior's decision to issue instructions in response to a request, and not to refer the problem further up the hierarchy to his superior amounts to a gamble on his part, since the superior then takes responsibility for the outcome of his subordinate's actions: he gets the credit if the results of the subordinate's actions are favorable, and the blame if it is otherwise.

Potential gambles - i.e. opportunities for personal gain by violating instructions, or exceptional situations which may appear to a subordinate to require his discretion rather than a literal application of the rules, will arise with greater or lesser frequency in all organizations. Whether employees find these opportunities profitable or not depends on the organization's incentive system. A useful special case is the "pure authority" incentive system, defined as one where no member of the organization other than its head takes any independent action, but all simply follow rules or orders from above. Equation (1) shows how such a system may be implemented. It suffices to raise p and f to sufficiently high levels that the expected gain to subordinates from all potential gambles, i.e. from independent action of any sort is negative. Consequently, all subordinates who are either risk-neutral or risk averse will be in equilibrium at the "corner" solution which corresponds to zero risk-taking, and will earn the riskless income Y_0.

To put it differently, in a system of pure authority everyone adheres slavishly to the rules; if an "exception" arises a subordinate will consult his superior as to the appropriate course of action. He in turn will consult his superior, and so on until the top of the hierarchy is reached. All decisions are made there; the orders are then passed down the pyramid to the lowest levels, where instructions are carried out.

Thus described, the pure authority model is not much different from the economist's standard caricature of internal allocation processes. To show why it represents a special case, we must first characterize the technology of an organization from a particular point of view.

In general, a technology may be defined as a set of procedures or operations for getting useful output from inputs. A task is a subset of these, assigned to an individual or a department (the user of the technology). A task is relatively routine, if the productivity of user discretion is relatively low, i.e., the appropriate procedures to be followed may be codified into a set of rules, such that relatively few "exceptions" arise in the course of production, i.e. relatively few situations arise where the user could obtain more output by independently choosing or modifying the procedures. The technique of auto assembly is an obvious example of a routine technology. Other technologies (medicine, social work) are much less routine: in these cases, procedures for getting the maximum output from inputs may not be specified without reference to the actual circumstances of production, and the marginal product of user discretion is relatively high[2/].

Now, if the technology of production is perfectly routine, so that there are no exceptional or ambiguous cases, the pure authority system cannot be improved upon as a system of organizational control. The manager of the organization translates this set of rules into the appropriate organ-

ganizational incentive system, simply by attaching the appropriate penalties and levels of monitoring such that the expected gain to subordinates from disobedience is always negative.

If, on the other hand, the technology of production is non-routine, the pure authority system suffers from two important defects. Firstly, the simple fact of "administrative overload" militates against its consistent use: in any organization of reasonable size the number of exceptional cases will be so large that the manager of the organization cannot possibly deal with them all himself.

The second defect is best illustrated with Tullock's (1965) model of hierarchical control. Tullock assumes that a certain amount of information is lost each time it passes through an additional level of hierarchy. It follows that the futher decision-making is removed from the level where decisions are carried out, the greater the amount of information distortion, or "control loss", to use Tullock's term. Hence, decisions made "at the top" will be based on distorted versions of what is going on at the bottom of the hierarchy, and the execution of commands at the bottom based on distorted versions of what was intended by those at the top.

For these reasons, managers must delegate authority to their subordinates. Indeed, a variety of rewards are used to encourage discretion, including bonuses, promotions, and non-pecuniary rewards such as titles, supervisors' praise and so on. Subordinates may of course get direct pleasure from the latter, but if not directly accompanied by increased renumeration, they surely often "signal" an increase in the probability that the subordinate will shortly receive a bonus or promotion. We ignore issues arising from the interaction of monetary and non-monetary rewards, and assume the simplest possible reward system. Subordinates are rewarded

who a) exercise their own discretion in cases where the rules do not apply, or are clearly inappropriate, and b) can later justify their actions to their superiors.[3/] We assume a fixed bonus (h) for each instance in which a subordinate exercises his own discretion in a way that pleases his superiors.

The incentive system facing subordinates in jobs with non-routine technologies therefore consists of the basic wage (Y_o), the level of monitoring (p), the sanction for discretion which is not justified (f) and the reward for justified discretion (h). How these different incentives operate, in terms of their effects on the way in which subordinates allocate their working time, is examined in the next section.

3. A Model of Subordinate Allocation Of Working Time

We assume that a subordinate allocates his working time to maximize the expected utility of final wealth over the length of his contract (T*). T* may be divided into categories of working time which yield different returns per unit of time:

$$T^* = t_a + t_h + t_g \tag{2}$$

Where t_a = the length of time where a subordinate does not act independently, but follows rules or instructions from above. (In the pure authority system $t_a = T^*$);

t_h = the length of time in which a subordinate exercises discretion which is "productive" from the organization's point of view, i.e., he acts in order to advance the purposes of the organization, in the expectation of a reward if such actions are noticed and approved by his superior;

t_g = the length of time in which a subordinate exercises "unproductive" discretion, from which he gains utility directly, i.e. without his actions being discovered by his superiors. t_g includes not only such things as

shirking and escaping discipline, but also exercising one's personal morality or personal preference, discretion or nepotism in policy decisions. The return to t_g may also be pecuniary, as in embezzlement, theft of organizational property, or the selling of organizational favors, such as accepting a bribe in exchange for the award of an organizational contract.

We assume that for each subordinate, the return to t_g may be summarized by an aggregate index of the opportunities for pursuing these activities available in his particular line of work. The monetary equivalent of this index is g, the average return per unit t_g. Since g is a monetary equivalent, it varies both over jobs (differences in opportunities) and over different persons in the same job if their net evaluation of the same opportunities differs.

For either kind of discretion, there is some probability that a subordinate who is discovered can justify his action to his superiors. The (subjective) probability of justification, z_i, is assumed for simplicity to be the same for both kinds of discretionary actions, but justifying productive discretion nets a subordinate the reward h, while "justifying" an "unproductive" descretion simply means that a subordinate is allowed to get away with it - he is not subjected to the sanction f.

Although t_h and t_g were distinguished above by the fact that t_g yields utility directly, while t_h does not, they are also distinguishable by potential productivity. Production discretion increases productivity whenever it is justifiable, unproductive discretion never does so. Since the probability of justification is constant, the average productivity of discretion is directly related to the fraction of discretionary time in t_h, i.e. to the fraction $\frac{t_h}{T}$.

In this model, then, each subordinate maximizes expected utility by allocating his working time among t_a, t_h, and t_g subject to the returns to each category of time, set by the organizational parameters Y_o, h, f, and p, and given his ability to justify discretionary action z_i, and the opportunities available in t_g as subsumed in g. If we define

$$T \equiv t_h + t_g$$

$$b \equiv \frac{t_h + t_g}{T^*} \equiv \frac{T}{T^*}$$

and

$$q \equiv \frac{t_h}{T}$$

then the subordinate's maximization problem may be expressed as his choice of an optimal level of risk taking (total discretionary time T) expressed as b, the fraction of total working time in which he acts on his own authority, and his selection of q, the optimal allocation of discretionary time among the risky "assets" t_h and t_g. 4/

A simple version of the subordinate's choice problem is expressed in the following model. Each subordinate maximizes

$$(3) \qquad EU(Y) = pz_i\, U(Y_o + hq\theta(b)) + p(1 - z_i)U(Y_o - f\theta(b)) + (1 - p)U(Y_o + g(1 - q)\theta(b))$$

subject to the constraints that

$$(4) \qquad 0 \leq q \leq 1$$

$$0 \leq b \leq 1$$

where U = the von Neumann-Morgenstern utility indicator; Y = monetary equivalent of final wealth over the time period T*; Y_o = the basic wage, expressed as the total basic wage received over T*; z_i is the subjective probability of justification, assumed constant over the time period T*. 5/

θ = the total number of discretionary actions over T*: $\theta \equiv \theta(b) \equiv kt$ where k is the number of different tasks perfomed per unit time. We assume for the present that k and T* are fixed, T* perhaps by bargaining, and k by the nature of the task. If k and T* are fixed, then since $\theta = kT$, $\frac{\theta}{T^*} = \frac{kT}{T^*}$ and the number of discretionary actions per unit time is uniquely related to the total number of discretionary actions over T*. The effects of variations in k and T* are dealt with subsequently (see footnote 14).

The first-order conditions for an interior maximum of (3) with respect to the choice variables q and b are [6/]

(5) $$\frac{h}{g'} = \frac{(1 - p)U'_g}{pz_i U'_h}$$

(6) $$pz_i U'_h hq + (1 - p)U'_g g'(1 - q) - p(1 - z_i)\, U'_f f' = 0$$

where U'_h = the marginal utility of income in state h, and a similar interpretation holds for U'_g and U'_f. Equation (5) shows the equilibrium allocation of discretionary time between the risky assets t_h and t_g. The term on the right-hand side is the marginal rate of substitution between t_h and t_g, and the term on the left hand side is the marginal rate of transformation. In equilibrium, they must be the same.

Equation (6) gives the equilibrium level of risk-taking, i.e. the allocation of time between riskless activity (t_a) and discretionary (risky) activity (T). Equation (6) states that discretionary activity will be carried to the point where the expected utility from the marginal hour in $T(\equiv T_h + t_g)$ equals the marginal utility of t_a.

Since the basic wage Y_o is fixed, the utility from the marginal hour in t_a is equal to zero, and consequently, discretionary activity will be carried to the point where its marginal expected yield (the left hand side of (6)) is equal to zero.

That the marginal utility of time is zero may appear counter-intuitive. However, it follows simply from our assumption that the basic wage is fixed,

and that subordinates receive no extra reward for performance of routine work (t_a). In any case, as in all portfolio choice models, it makes no difference to the properties of the model whether the riskless asset is defined to have a positive or zero yield. Thus, t_a could be defined to have a positive return, in which case, in equilibrium, the marginal expected utility of T would still equal that of t_a, but this number would be positive rather than zero. The model would be more complicated, but its comparative static properties would be unchanged.

An important implication of (6) is that the level of discretion is chosen by employees, and not fixed by the nature of the task or by the employer. For example, it might be thought that b is set by the nature of the task, since, if the task is non-routine, employees will "need" to use more discretion than they would for routine tasks. However, as shown elsewhere, (pages 5 and 17-20), the nature of the task sets the productivity of discretion, not its level; ceteris paribus, when that productivity is relatively high, the employer interested in efficiency will encourage discretion by selecting appropriate values of h, f, and p. Subordinates choose the level of discretion, as in equation (6) given its "price" as determined by the values of these incentives. They can be "forced" neither to obey rules nor to exercise discretion.

Indeed, in some circumstances, employees deliberately choose not to exercise discretion, despite the obvious productivity of doing so, as in the well-known phenomenon of "work to rule", or in an example given by Bendix, in the case of inmates of concentration camps, who, assigned to work in factories, sabotaged production up to 80% by the simple tactic of consistently asking for instructions on what to do next[7].

I now turn to the comparative statics of the model, i.e., to the effects of changes in the incentives under the manager's control (h, f, and p)

on a representative employee's optimal choices of q and b. The results are presented and interpreted in the text; mathematical derivations may be found in the appropriate footnotes.

By totally differentiating equation (6) it may be shown[8/] that an increase in f, the sanction applied to any subordinate who fails to justify an action taken on his own authority, reduces total discretionary time T. T must decrease, since the increase in f reduces the expected returns to both kinds of discretion (t_h and t_g). Totally differentiating euqtion (5) with respect to f leaves q unchanged. The reason is that, since the same sanction is imposed for both unjustified productive and unjustified unproductive discretion, and since z_i is also the same for both t_h and t_g, the expected returns to t_h and t_g fall by the same proportion when f is increased. t_h and t_g therefore decline by the same amount, and q is unchanged.

If the probability of justification were different for t_h and t_g, the strong conclusion that $\frac{\partial q}{\partial f} = 0$ would not hold; if z_i were higher for t_h, it can be shown that $\frac{\partial q}{\partial f} > 0$, and similarly $\frac{\partial q}{\partial f} < 0$ if z_i were higher for t_g. In all cases, however, an increase in f would lower both t_h and t_g. This is the chief limitation on the use of sanctions to control subordinate behavior: heavy sanctions deter subordinates from independent action of any sort, and not just from actions which are against the interests of the organization.[9/]

Our model assumes a unique sanction. If multiple sanctions were available, the organization would not be limited in its ability to penalize certain kinds of behavior which are both clearly unjustifiable, and which are easily distinugished from productive discretion, such as theft, and in some, but not all cases, shirking, and sanctions would be relatively high in these cases.

However, in the more interesting areas of bureaucratic decision-making, productive and unproductive discretion are not so easily distinguished, nor can employees easily anticipate whether their discretionary actions will be ruled justifiable or not. Since the effects of discretion on output are not directly measurable,[10/] the true "facts" are inherently ambiguous, and there is always room for argumentation and error over matters such as what constitutes an exceptional case, whether the employee's actions were indeed correct under the circumstances or not, etc. Consequently, even if the manager were to specify different sanctions for unjustified productive and unjustified unproductive discretion, errors in their application would still produce the result that larger sanctions reduce both t_h and t_g Only if it were costless to distinguish t_h from t_g with perfect certainty, would this result not hold.

Both the effects of a change in h, the reward to productive discretion which is justified, and in g', the marginal valuation of unproductive discretionary time, on q and b may be shown to depend on the relative magnitude of opposing wealth and substitution effects.[11/] With respect to h, however, we demonstrate in section 4 that substitution effects must be dominant on the average, i.e., that an increase in h raises the average levels of both q and b for any group of subordinates subject to a common control system. (see page 25).

Obviously, subordinates who are relatively able, or who are simply more clever and ingenious than other employees at presenting a case that their discretionary actions contribute to organizational goals, and whose subjective probability of justifying the discretionary use of their time (z_i) is higher, exercise a relatively large amount of discretion[12/].Less obviously such employees do in fact devote a higher fraction of their discretionary time to productive uses, i.e. $\frac{\partial q}{\partial z_i} > 0$. Note that although these

employees exercise more discretion that others, they do not necessarily take more risk, since both productive and unproductive discretion are simply less risky alternatives for them than for other employees. More able employees therefore earn higher average wages than others, even if their basic wage Y_o is the same, via the effect of their higher z_i in increasing q and b and therefore in increasing total wages Y.

An increase in the level of monitoring or supervision (p), as approximated, perhaps, by the span of control, tends to increase productive discretion (t_h) and to diminish unproductive discretion (t_g).[13/] Conversely, employees in jobs where the level of monitoring is relatively low, exercise relatively little productive discretion. Since the monetary income of subordinates who tend to be ignored by their superiors therefore tends to be relatively low, they compensate by taking a relatively large fraction of their income in non-monetary and non-organizational forms by selecting relatively high levels of t_g. Finally, total discretion is inversely related to p. The reason is that an increase in p tends to lower the return to t_g by more than it raises the return to t_h. Hence, differentiating equation (6) with respect to p yields $\frac{\partial b}{\partial p} < 0$.

This completes the analysis of the effects of changes in p, h, and f, on subordinates' allocation of time[14/]. The next section focusses on the manager's optimization problem, which takes into account the costs of the different instruments of control, as well as their effects on the behavior of subordinates.

4. The optimal incentive system

(a) The Manager's objective function

The objective which we impute to the manager of the control system - hencefore, the "manager" for simplicity - is to choose the efficient control

system, i.e. that system which maximizes the difference between the total value of output produced by a given number of production workers with the aid of control inputs and the total costs of control inputs. For private firms this is equivalent to ordinary profit-maximization. The output or services of public bureaus however, are typically not sold, and their managers cannot be assumed to maximize profits. Citizens do place a value on public services, however, and we view public bureaucrats as the subordinates of elected politicians who supply goods and services to citizens in exchange for political support.

If political competition is perfect, then elected politicians are forced to supply public services efficiently, and this implies they must choose efficient control systems, or lose office[15, 16/]. If competition is weak, e.g., because of barriers to entry in politics, then elected politicians, have some degree of freedom [17/]. to implement other control systems. Why they would choose to exercise their preferences by choosing an inefficient control system is, however, unclear[18/]. But monopolistic politicians might allow the control system to degenerate if they take their monopoly rents in the pursuit of the "quiet life" as could, of course, monopolistic private firms. The rationale for our assumption that managers of either public bureaus or private firms choose the efficient control system rests solely on the strength of competitive forces; where such forces are weak in the private sector, it is possible for the managers in firms to pursue objectives other than profit-maximization; some of these objectives have been investigated extensively in the literature[19/].

However, our formulation is sufficiently transparent that the consequences of assuming that control mangers have different objectives are often quite obvious, and will be indicated briefly in the subsequent discussion.

Managers of private firms therefore choose efficient control systems

to maximize profits, and elected politicians choose efficient control systems to maximize the probability of their re-election. Given this maximand (V), the model to be developed predicts the equilibrium values of p, h, and f as a function of basic organization characteristics, such as technology and factor complementarity, the costs of control personnel, the size of the organization and other variables. Changes in these exogenous variables cause efficiency-minded organization managers to select different sets of values of p, h, and f, and these changes induce subordinates in turn to change their optimal allocation of time. The analysis therefore predicts the equilibrium values of total discretion (T), and the allocation of discretionary time between productive and unproductive discretion (q) as well as values of p, h, and f, as functions of more basic characteristics of organizations.

To proceed with the basic model, we first state the managers' objective function, which incorporates the costs of control devices as well as their benefits.

The managers' objective is to select values of p, h, and f to maximize

$$(7) \qquad V = G(qT) - H_1(T) - H_2((1 - q)T) - C(p, T) - pT(q\bar{z}_i h - (1 - \bar{z}_i)f)$$

where the function G is an indicator of the degree of routinization of the technology of work; H_1 and H_2 are, respectively, the damages due to external effects imposed on complementary factors when subordinates manage their own working time, and the "private" damages (value of lost output) due to unproductive discretion; C is the control cost function; and the term $pT(q\bar{z}_i h - (1 - \bar{z}_i)f)$ represents the costs of rewards paid out minus the revenue received from sanctions imposed. In that term, $\bar{z}_i$ is the average probability of justification of the group of subordinates to whom the control system applies, and the total number of discretionary actions is

represented by T instead of $\theta = kT$ to simplify the notation.

We now discuss each of these relationships in detail. The function $G = G(qT)$, or equivalently, $G = G(t_h)$, characterizes the degree of routinization of the organization's technology of production as the value of productive discretion on the part of subordinates. For all non-routine technologies, therefore $G' > 0$. The marginal value of productive discretion (G') is the indicator of technology, i.e. G' varies directly with the degree to which the technology of production is non-routine. The basic determinants of G' are the organization's division of labor, the skills (human capital) of its employees, and the accumulated experience of the organization in using the technology. Each of these variables is discussed in turn.

The first two propositions are that the tasks of employees will tend to be more routine, the more extensive the organization's division of labor and the lower the average human capital of employees. The basic characteristic of the division of labor, as exemplified in Adam Smith's famous pin-making factory, is the subdivision of tasks into a series of narrow, repetitive sub-tasks. This subdivision obviously facilitates a more complete specification of the procedures to be followed in the performance of each sub-task than would be possible for the whole operation, and hence, G' tends to be inversely related to the organization's division of labour.

The effect of changes in the division of labor on G' should be distinguished from the effect due to accompanying changes in the average skills (human capital) of employees. Ceteris paribus, an increase in the level of human capital, tends to increase G', since the productivity of discretion obviously depends on the skill with which subordinates exercise it[20/].

The factoring of production into relatively narrow, repetitive, and simple operations tends to permit the employment of relatively unskilled personnel, implying a reduction in skill levels as the division of labor increases. On the other hand, specialization according to comparative advantage, which plays no role in Smith's discussion, might sometimes imply the opposite, i.e. that the average skill levels of employees increase with the division of labor. Consequently, there is a priori no necessary correlation between average skill levels and the extent of an organization's division of labor.

Empirically, the division of labor may be approximated by some index of the number of different tasks into which production is broken down, such as the number of different job titles within an organization[21] and skill levels or human capital by average wage rates. G' will be negatively related to the number of job titles, and positively related to the average wages of employees.

The second set of factors which affect G' all relate to the extent of an organization's experience with a technology. In particular we shall show that G' is positively related to an organization's "technical progressiveness" (rate of increase in productivity)[22], and therefore, like technical progressiveness, G' tends to be inversely related to the age of the organization, the age of its capital stock, and the age of its industry.

The notion of technical progressiveness is based on the "progress function",i.e. the finding that an organization's productivity tends to increase with experience, measured either by age or cumulative output.[23] This increase in productivity is usually interpreted as a learning effect, although the process by which organizations "learn" is seldom elaborated upon, and indeed, the sense in which organizations, as opposed to individuals, may be said to "learn" is certainly unclear.

In our framework, it is easy to see how such improvements in productivity come about as experience with a technology is accumulated. As the

firm ages, or produces more output, exceptions (unanticipated contingencies) occur, and standard procedures are either adapted or new contingent procedures devised to deal with them. Useful knowledge is accumulated and technical progress achieved if those adaptations or new procedures may be used again in the future. This will be the case, provided only that some fraction of these contingencies do tend to recur in future periods, and that curcumstances do not change so rapidly that the procedures invented are not useful for subsequent applications.

The larger the number of contingencies which are not genuine "surprises", but tend to recur with stochastic regularity, and the slower the rate of "depreciation" of the procedures devised to deal with them, the faster the rate at which knowledge can be accumulated.

The organization, and not merely the individuals in it, "learns" when this knowledge is not merely accumulated by employees[24/] but is institutionalized in the form of either standard or contingent rules and procedures, and incentives to use these rules built into the control system in the manner discussed previously. In this way, organizational rules may evolve over time to encompass what were once exceptional contingencies which were either inappropriately handled by existing rules, or required discretionary action on the part of subordinates.

This process partly explains why an organization's productivity tends to increase with time or cumulative output. It also implies that, in the absence of new stimuli, such as new investment[25/] the scope for discretion tends to diminish over time, as efficient ways of dealing with the more important problems are discovered and institutionalized in organizational rules[26/].

Consequently, in the absence of new stimuli, the marginal product of discretion will tend to diminish over time, and this diminution will be accompanied by a fall in the rate of increase of productivity. The rate of increase of productivity, and the marginal product of discretion, will also

both tend to be related to the age of the industry (as a proxy for the "start" of the technical progress function[27]) and to the age of the organization's capital stock, since changes in the capital stock both raise new opportunities for organizational "learning" and depreciates the knowledge built around the old capital stock.

An important implication of this argument is that the degree to which the technology is routine is unrelated to the absolute level of productivity. It is related to the rate at which the organization's technology is improving and therefore it may be either high or low in organizations where the technology works well in an absolute sense. This explains why organization theorists [e.g. Woodward, (1965)] often find similarities between the control systems used for production processes which are the "most" and "least" advanced from a technical point of view. A routine control system is indeed entirely appropriate to technology of production which is relatively primitive and not well understood, (education is a good example) if the expected increase in productivity from implementing a less routine system is small.

It is clear from equation (7) that the choice of the optimal control system does not depend simply on technology. In particular, and apart from the direct costs of control instruments, there are two sources of damages which result when authority is delegated to subordinates to manage their own working time, the magnitudes of which will be taken into account by the manager in choosing the efficient control system.

The first source of damages, H_1, is due to intra-organizational external effects imposed on complementary factors (lack of coordination) when subordinates are free to manage their own working time. H_1 depends on the level of total discretionary time, T, i.e. $H_1 = H_1(T)$, where $\frac{\partial H_1}{\partial T} \equiv H_{1T} > 0$. The reasons for this source of loss have been adumbrated at length elsewhere, and briefly summarized in footnote 1 above. Although that argument is couched in

terms of the choice between market and internal organization, the analysis applies equally well to the choice, within a hierarchy, of more vs. less routine methods of control: to the extent that subordinates are free to choose their own procedures, their decisions will necessarily impose external damages on complementary intermediate outputs of other subordinates. One way to reduce these damages, i.e. to coodinate the activities of employees, is to restrict their discretion. Hence, $H_{IT} > 0$. The size of the damages will vary directly with the degree of complementarity of the different factors and with the range of factors (size of organization) which incur damages. The formulation used in equation (7) does not distinguish between these two, and we shall refer simply to the total damages H_1.

The damages from unproductive discretion appear as $H_2(t_g)$ in equation (7) where $H'_2 > 0$. H_2 therefore represents the foregone value of subordinate's working time, plus any additional private damages resulting from unproductive discretion. Of course, unproductive discretion will also yield external damages, reflected in the value of final output produced with the aid of complementary factors. This source of damages has already been included in H_1 and is not counted again in H_2.

The cost function $C = C(p, T)$ comprises the direct costs of operating the control system, i.e., the costs of monitoring and evaluating subordinates' performance. In organizations large enough to take advantage of it, the principle of comparative advantage dictates that these tasks will be specialized, i.e., performed by different persons or departments, and therefore a third control function - that of communicating the relevant information between monitors and evaluators - must be included.

Each of these costs will vary among different types of organizations; for any given organization, however, we expect that, holding the number of production workers constant, the costs of monitoring varies directly

with its level, as measured by the probability that a given subordinate's actions will be discovered, i.e. $C_p > 0$.

We also expect that $C_T > 0$: if the probability of being discovered is held constant, the costs of communication and decision-making increase directly with T, the number of discretionary actions which require evaluation if discovered. In addition, we expect the costs of monitoring to increase as T increase: when T is relatively low, subordinates perform the same (routine) operation most of the time, and economies of scale in monitoring may be achieved by sampling, since any one subordinate's activities correspond more closely to a random sample of both what other subordinates are doing, and of what he himself is doing in periods when his actions are not directly observed. These scale economies are forfeited as discretion increases, and therefore the costs of monitoring increase directly with the average level of subordinates' discretion.

The effects of information coordination and control costs on the efficient control system may be summed up briefly. All of these costs tend to reduce the efficient level of discretion below that which would be predicted on the basis of "technological" considerations alone. The efficient control system therefore always tends to appear "over routinized" from the subordinate's point of view, i.e. subordinates will always be able to think of actions which appear to them to be more efficient than the standard and prescribed ways of doing things, and which they believe they could justify to their superiors, if only sufficient incentives were provided to do this.

To the extent, however, that relatively weak incentives for the productive exercise of discretion reflect the extra costs of information, control and coordination incurred as discretion increases, this "bureaucratization" of production is not inefficient, but correctly modifies the inappropriate "signals" which employees receive from their immediate environment.

(b) Optimality conditions

The first-order conditions for an interior maximum of (7) with respect to the choice variables p, h, and f may be written as follows:

$$(8) \qquad T_f(G'q - H_{IT} - H'_2(1 - q) - C_T) = T_f(p(q\bar{z}_i h - (1 - \bar{z}_i)f) - p(1 - \bar{z}_i)T$$

$$(9) \qquad T_h(G'q - H_{IT} - H'_2(1 - q) - C_T) + q_h(G' + H'_2)T = T_h(p(q\bar{z}_i h - (1 - \bar{z}_i)f) + p\bar{z}_i(q + hq_h)T$$

$$(10) \qquad T_p(G'q - H_{IT} - H'_2(1 - q) - C_T) + q_p(G' + H'_2)T - C_p = T_p(p(q\bar{z}_i h - (1 - \bar{z}_i)f) + (\frac{AR}{p} + pq_p\bar{z}_i h)T$$

These first-order conditions may be immediately interpreted as stating that each instrument will be used to the point where its marginal benefits are just equal to its marginal costs. To simplify the notation, let

$V_o \equiv G'q - H_{IT} - H'_2(1 - q) - C_T$; $V_1 \equiv G' + H'_2$; $AR \equiv p(q\bar{z}_i h - (1 - \bar{z}_i)f)$

V_o is the marginal net gain in the value of output from an increase in T, i.e. the increase in the value of output minus the increase in control costs; V_1 is the marginal gain from an increase in q; AR = average net rewards.

If we divide each of the first-order conditions (8), (9), and (10) by T_f, T_h, and T_p respectively, the first-order conditions may be alternatively expressed in terms of the marginal gains and costs of total discretionary time T. The first order conditions may therefore be re-written as follows:

$$(8)' \qquad V_0 = AR(1 + \frac{1}{E_f^T/E_f^{AR}})$$

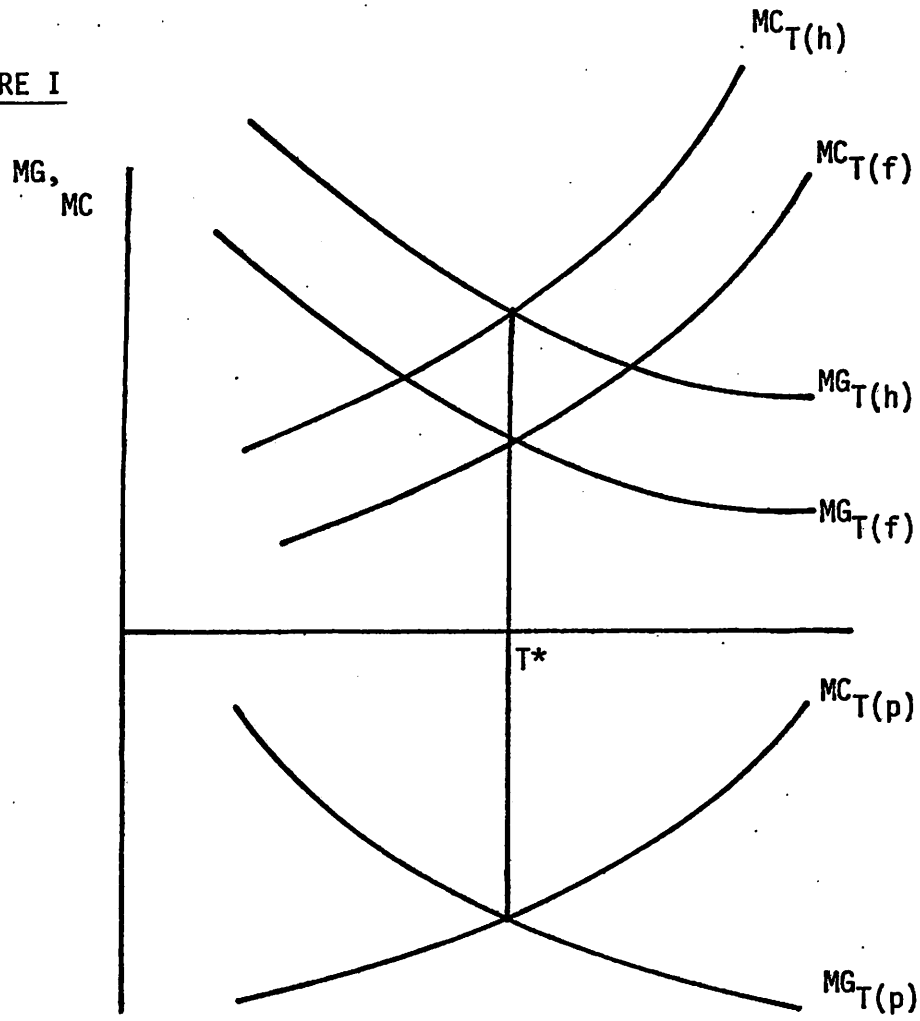
$$(9)' \qquad V_o + V_1 \frac{E_h^q}{E_h^T} q = AR(1 + \frac{1}{E_h^T/E_h^{AR}})$$

$$(10)' \qquad V_o + V_1 \frac{E_p^q}{E_p^T} q - \frac{C_p}{T_p} = AR(1 + \frac{1}{E_p^T/E_p^{AR}})$$

where $E_f^T = T_f \frac{f}{T}$ and a similar interpretation holds for E_h^T and E_p^T; $E_f^{AR} = AR_f \frac{f}{AR}$, and a similar interpretation holds for E_h^{AR} and E_p^{AR}.

Equations (8)', and (9)' and (10)' state the marginal benefits and costs of each instrument in terms of T, which is advantageous in that the curves representing the costs and benefits of all three instruments may be shown on a single diagram (see figure 1).

FIGURE I



These first-order conditions may each be interpreted as stating the following optimality condition for each instrument: an instrument will be used to the point where the marginal gain from the change in discretion (T) plus the change in the average productivity of discretion (q) achieved by its use are just equal to the marginal costs of the additional discretion "purchased" by the use of that instrument.

Second-order conditions are consistent with, but do not uniquely imply, decreasing marginal gains and increasing marginal costs from discretion ($G'' < 0$, and H_{TT}, $C_{TT} > 0$) as drawn in the figure.[28/]

Before turning to the comparative statics of the model, note that equations 8'-10' imply that, in the relevant regions, E_h^T and E_h^q are both greater than zero. In the model presented in section 3, the signs of T_h and q_h depended on the relative magnitudes of opposing wealth and substitution effects. It was shown, however, that T_h and q_h will always have the same sign, whether positive or negative. It follows that if E_h^T is negative, E_h^q must be negative also. But if both E_h^T and E_h^q were negative for a group of subordinates subject to a common reward system, the manager could increase both q and T by decreasing the size of the bonus h. Provided only that the increase in T is desirable ($V_o > 0$), both the increase in q and the increase in T increase the value of organizational output. At the same time, the total costs of net rewards will be lower, since $AR_h > 0$; the manager will therefore continue to decrease h until a region is reached where E_h^T and E_h^q are > 0. The region where E_h^T, $E_h^q < 0$ is uneconomic; no rational manager will continue to operate there.

(c) Comparative Statics

This section presents a number of empirical implications of the model. Again, the text is reserved for presenting and interpreting the results

and derivations may be found in the appropriate footnotes. Only the unambiguous results are reported in the text. The exogenous variables considered here are changes in the damages due to lack of coordination (H_1), the damages from unproductive discretion (H_2), the costs of control, the size of the organization, and in the set of variables which affect G'- the age of the firm, age of its capital stock, and industry, average human capital of employees (wages), and the division of labor. All of these variables are measurable, with the possible exception of H'_2, for which I have been unable to suggest an appropriate proxy, and the elasticity of substituion, (which enters H_1) for which the difficulties of measurement are well known.

More difficult problems of measurement are encountered for the dependent variables. The theory predicts values of p, h, and f, discretion (T, t_h) and the fraction of time devoted to productive discretion (q) as functions of the exogenous variables. h and f are of course, theoretical constructs, the institutional counterparts of which are such things as merit pay, bonuses, raises, promotions and demotions, outright firings, etc. Direct measurement of the average or marginal values of rewards and sanctions is no doubt both possible and exceedingly difficult. The level of monitoring (p), however, may be directly approximated by the span of control whenever the productivity of control personnel may be assumed constant, or, if the ratio of control personnel to direct labor is constant, by the ratio of "administrative" personnel to direct labor. Data for both these measures are readily available and extensively used by organization theorists and sociologists [29/].

The heart of the model, however, is its predictions with respect to the level of discretion, and adequate tests of the theory can be performed if reasonable proxies for this variable can be devised. There have indeed been numerous attempts by sociologists and organization theorits to directly measure one variant or another of the variable which we term discretion [30/].

The major problem with these studies is that they are all based on interview data ("how routine is your job?") and hence unreliable.

An empirical proxy for discretion based on objective data may be arrived at, based on the following reasoning. When values of the incentives h, f and p encourage relatively large amount of discretion, some employees will be successful in exercising it and therefore rewarded during any particular time period. The discretionary actions of others will not be discovered while still others will be discovered and fail to justify what they do, and be subjected to sanctions. Consequently, high average levels of discretion will be accompanied by a relatively high dispersion in the earnings of employees. Conversely, relatively low levels of discretion are reflected in a relatively "flat" earnings structure, i.e. one in which all employees at a particular level tend to earn the same income.

Therefore, a measure of discretion which can be used to compare average discretion at different organizational levels, or at the same organizational level in different organizations, is the coefficient of variation of organizational income (i.e., excluding income from other sources):

$$\frac{\sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}}{\bar{Y}}$$

where Y_i is the organizational income of the i'th employee, $\bar{Y}$ the mean organizational income, and $\underline{n}$ the number of employees at the level of the organization.

To compare discretion between organizations with different number of levels, one needs an index of the average level of discretion in an organization. This is given by

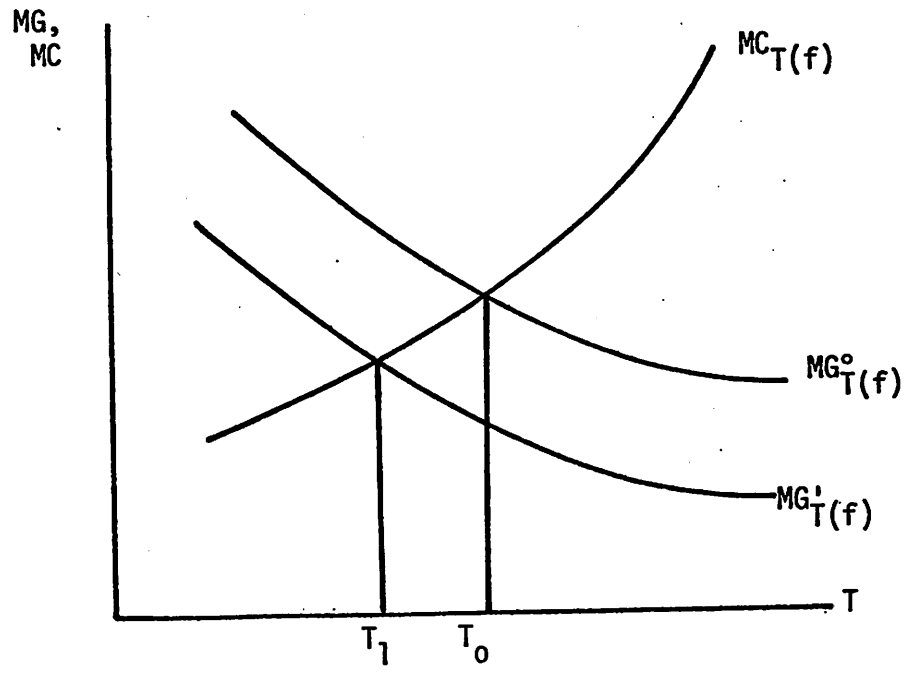
$$\frac{\sum_{j=1}^{m} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})}}{m\,\bar{Y}}$$

where m is the number of organizational levels

Note that one cannot use the coefficient of variation of an individual's earnings to measure his discretion, since this will reflect variations in the base wage paid to all employees due to changes in demand for the firm's product, factor scarcities, etc. Dispersion in wages paid to different employees at the same level, however, reflects differences in their perceived performance, and such differences, whether due to differences in abilities, luck, or attitudes towards risk on the part of employees, can only arise via their exercise of discretion [31/].

I now turn to the comparative statics of the model. H_{1T}, the marginal external damages due to subordinate discretionary activity would increase if the complementarity between the tasks of employees or the size of the organization increases. The increase in H_{1T} would induce managers to increase f, decrease h and increase p in order to discourage discretion. [32/] Each of these changes would indeed induce subordinates in turn to decrease their allocation of time to discretionary activity, as shown in Section 3, and hence the value of T decreases unambiguously. These results are depicted in figure II, where the increase in H_{1T} is shown to decrease $MG_{T(f)}$, and hence the optimal level of T decreases from T_0 to T_1, implying an increase in optimal f. Since the increase in H_{1T} also decreases $MG_{T(h)}$ and $MG_{T(p)}$, figure II can also be used to illustrate these effects, merely by re-labelling $MG_{T(f)}$ as $MG_{T(h)}$ or $MG_{T(p)}$, and $MC_{T(f)}$ as $MC_{T(h)}$ or $MC_{T(p)}$

FIGURE II



The case of a change in H_{1T} can also be used to illustrate the consequences of dropping our assumption that the manager always chooses the efficient control system. If competition is weak, then the manager has some discretion to implement his personal preferences in designing the control system. If a manager is addicted to personal power and dictatorial control, this implies an increase in the damages (to him) from discretion. If this managerial preference were incorporated into the model, the consequent changes in the values of p, h and f would be identical to those just discussed for an exogenous increase in H_{1T}. If managers have other personal objectives, the consequences of implementing them could be illustrated in the same way, i.e., by adapting the relevant coefficient in equation (7).

A fall in the marginal product of discretion induces managers to increase f, and to decrease h and p.[33/] Each of these changes reinforces each other in inducing employees to reduce productive discretion (t_h), and to reduce the fraction of total discretion which is productive (q).

Total discretion will also fall as long as the combined effects of the changes in h and f dominate that of the change in p. And indeed, it can be shown that E_p^T tends to be "small" relative to E_f^T alone, and therefore, a fortiori, the change in p is unlikely to dominate that of the combined effects of the changes in h and f.[34/] Accordingly, we predict that firms will tend to be more routine, as reflected in a smaller dispersion among the incomes of their employees, the older the firm, the older its capital stock, the lower the average wages of its employees, and the older the industry of which it is a member.

An increase in the division of labor within an organization, ceteris paribus, tends to diminish G', leading to a reduction in discretion. The division of labor also tends to be accompanied by an increase in the scale of operations, and perhaps also by an increase in the degree of complementarity. For both these reasons, the marginal external damages due to discretion, H_{1T}, will tend to be larger, the more extensive the division of labor. Both the independent effect of the increase in size(discussed later in this section) and the effect of the increase in H_{1T}, would induce managers to increase optimal f, decrease optimal h, and increase optimal p, implying a decrease in the average level of discretion T. These effects reinforce that of the effect of the division of labor through technology, implying an inverse correlation between the extent of an organization's division of labor, and the average level of discretion of its employees.

Although no systematic evidence has ever been collected on these

hypotheses, some scattered observations may be made. First, the theory predicts a positive correlation between the productivity of discretion and the level of monitoring (p). Since p tends to be inversely related to the span of control, this prediction is consistent with the fact that the span of control of "first-line" supervisors tends to be larger than that at higher levels, and more generally tends to fall as one moves up the hierarchy.[35/] The analysis also predicts that the dispersion of incomes will tend to be relatively larger among employees at higher organizational levels than among those at lower levels and this is certainly consistent with everyday observation. Thirdly, the predicted inverse correlation between span of control and average wages is consistent with evidence that the span of control in "professional" organizations tends to be smaller than in other organizations of the same size, i.e., the former tend to be relatively tall, narrow hierarchies.[36/] Finally, and of particular interest, is the data collected by Mansfield [1962] on the age of different industries, which strongly shows that the extent of firm mobility within an industry, i.e., the extent to which firms change their relative positions in the industry's size distribution, is inversely related to its age. Our analysis predicts that older industries will tend to be populated by relatively routine firms, and this is consistent with Mansfield's evidence of lack of mobility in those industries.

The value of damages from theft, shirking, lack of discipline and other forms of unproductive discretion, H'_2, might vary between organizations, either because opportunities for employees to enrich themselves at the expense of the organization vary, (e.g. employees who award contracts to private firms can usually do more damage than those who are limited to stealing pencils and memo pads), or because malfeasance has different consequences in different organizations. Within an organization, H'_2, is undoubtedly correlated with an employee's height in the organizational

hierarchy. An increase in H'_2 induces managers to increase the optimal values of h, f, and p to reduce the level of unproductive discretion and increase q.[37] One interesting implication of this analysis is the use of relatively high rewards to accomplish this end, as well as close monitoring and relatively high sanctions which discourage malfeasance directly.

We now consider changes in the control cost function, C(p, T). The costs of monitoring and evaluating subordinates' performance would increase if either the wages of monitors or evaluators were to increase, as perhaps would be the case where their tasks were more difficult, and hence required the employment of more specialized personnel. In such cases, both C_p and C_T would increase, causing managers to decrease optimal h and increase optimal f in order to discourage discretionary activity.[38] Optimal p may either increase or decrease. Discretion will certainly fall if p increases, and it will fall even if p decreases, as long as the decrease in p is not so large as to dominate the effects on discretion of the changes in f and h.

The effects on control costs and on discretion of a change in the size of the firm are particularly interesting and important. There are two effects: firstly, the costs per employee of controlling and coordinating discretionary activity (H_{1T} and C_T) increase directly with firm size. An increase in the number of employees means that any given level of subordinate discretion implies an increase in the absolute number of cases which must be communicated to and evaluated by the manager. The quality of the manager's decisions would be reduced because of the administrative overload factor, and because of the increase in "control loss" incurred as the number of hierarchical levels through which the information must be passed increases. Secondly, the average costs of monitoring routine activity decrease, due to economies of scale in sampling obtained as the size of the organization increases: the larger the number of employees performing the same routine operation, the lower

the costs per man of obtaining information on any characteristic of their activities to any specified degree of accuracy.

The first of these changes implies an increase in the rate at which the marginal costs of controlling and coordinating discretionary actions increases with the average level of discretion, and therefore an increase in C_T and h_{1T}. The effect of the second change is more subtle. Scale economies reduce the average costs of monitoring the routine activities of subordinates, but leave the cost of monitoring discretionary actions unchanged, since sampling only yields information concerning characteristics of subordinate performance which subordinates have in common, i.e., routine characteristics. Consequently, the effect of sampling economies is also to increase C_T, the rate at which total monitoring costs increase as discretion increases. C_T increases because allowing discretion implies foregoing the cost savings achievable by sampling, and these cost savings foregone are larger, the larger the size of the organization.

Both effects therefore reinforce each other in increasing the relative costs of controlling and coordinating discretion as size increases. Hence discretion must fall, via an increase in f, decrease in h, and possibly an increase in p. 39/

This change in relative costs as the size of the organization increases explains why the terms "large" and "bureaucratic" are practically synonomous descriptions of an organization. Most of the "evils" of bureaucracy - stifling rules and regulations, lack of innovation and capacity for flexibility, red tape, boring and repetitious work, lack of responsibility, endless delays in decision-making, etc. - clearly flow from this incentive for production to be increasingly routinized as the size of the organization increases.

Do these classic diseconomies of scale also provide a limit to the

size of the firm? It has recently been held that they do not, essentially because these diseconomies may be more than compensated for by the informational economies of scale obtainable at larger size.[40] [see, e.g. R. Wilson, (1975)]. Our classification of working time into discretionary and non-discretionary time shows why the classical argument, i.e., the argument that firm size is limited by increasing difficulties of coordination and control is nevertheless correct. The reason is that informational scale economies are only achievable in the control of routine operations. The costs of coordinating and controlling discretionary activity (H_{1T} and C_T) therefore continue to increase with size. Moreover, informational scale economies are undoubtedly diminishing with size,[41] while, because of the cumulative nature of the distortion in information implied in the control loss process, the diseconomies of scale in controlling discretion probably tend to increase with size.

Now, as size increases, firms do react to the increasing costs of controlling discretion and the falling costs of controlling routine behavior by reducing the incentives to employees to exercise discretion. But they cannot avoid ultimately increasing average costs by doing so, as long as the marginal productivity of discretion is diminishing ($G'' < 0$), which is surely the case. For then, as the firm becomes more and more routine with increasing size the loss in productivity from reduced discretion becomes larger and larger.

Consequently, as the firm increases in size, a point must be reached where either the increasing average costs of controlling discretionary activity or the increasing losses in productivity from reduced discretion must overtake the falling average costs of controlling routine activity, and this point is the limit to the size of the firm.[42]

Alchian, A., "Reliability of Progress Curves in Airframe Production", Econometrica, vol. 31, no. 4, 1963: 679-93.

Alchian, A., and Demsetz, M., "Production, Information Costs and Economic Organization", American Economic Review, 62 (1972)

Arrow, K.J., "The Economic Implications of Learning by Doing" Review of Economic Studies, Vol. 29, 1962: 155-73

Arrow, K.J., Essays in the Theory of Risk-Bearing. Chicago: Markham, 1971.

Baumol, W.J., Business Behavior, Value and Growth. New York: Little Brown, 1959.

Becker, G.S., "Competition and Democracy", Journal of Law and Economics 1 (1958): 105-9, reprinted in G.S. Becker, The Economic Approach to Human Behavior. Chicago: University of Chicago Press, 1976.

Becker, G.S., "Crime and Punishment: an Economic Approach", Journal of Political Economy, 76, no. 2 (March/ April 1960): 169-217

Becker, G.S., and Stigler, G.S., "Law Enforcement, Malfeasance, and Compensation of Enforcers", Journal of Legal Studies 3, no. 1 (January) (1975): 1-18

Bendix, R., Work and Authority in Industry. Berkeley, University of California Press, 1974.

Blau, P., and Schoenherr, P., The Structure of Organizations. New York: Basic Books, 1971

Breton, A., The Economic Theory of Representative Government. Chicago: Aldine, 1974.

Breton, A., and Wintrobe, R., "The Equilibrium Size of a Budget-Maximizing Bureau", Journal of Political Economy, vol. 83, no. 1, 1975.

Burns, T., and G.M. Stalker, The Management of Innovation London: Tavistock, 1961.

Crozier, M., The Bureaucratic Phenomenon. London: Tavistock, 1969.

Ehrlich, I., "Participation in Illegitimate Activities: A Theoretical and Empirical Investigation" Journal of Political Economy May/June, 1973.

Hage, G., and Aiken, M., "Routine Technology, Social Structure and Organizational Goals", Administrative Science Quarterly, 1969

Hirsch, W., Firm Progress Ratios, Econometrica, vol. 24, (1956): 136-43.

Hickson, D., "A Convergence in Organization Theory", Administrative Science Quarterly, September, 1966.

Kennedy, C., and Thirlwall, A.P., "Surveys in Applied Economics: Technical Progress", Economic Journal, March, 1972.

Mansfield, E., "Entry, Gibrat's Law, Innovation, and the Growth of Firms", American Economic REview, vol. 52, 1952: 1023-51.

Marris, R., The Economic Theory of 'Managerial' Capitalism, London, MacMillan, 1967.

Meyer, M., "Expertness and the Span of Control", American Sociological Review, vol. 33, no. 6 (1968).

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

Perrow, C., "A Framework for Comparative Organizational Analysis", American Sociological Review, vol. 32, no. 2, (1967): 194-208.

Pondy, L., "Effects of Size, Complexity and Ownership on Administrative Intensity", Administrative Science Quarterly, vol. 14, no. 1, 1969: 47-61.

Presthus, R., "Toward a Theory of Organizational Behavior", Administrative Science Quarterly, June, 1958.

Rosen, S. "Learning By Experience as Joint Production", Quarterly Journal of Economics, 1972

Rubin, P. "The Expansion of Firms", Journal of Political Economy, vol. 81, no. 4, 1973: 936-950.

Simon, H., The New Science of Management Decision. New York: Harper, 1960.

Stigler, G., "Economic Competition and Political Competition", Public Choice. Vol. 13, 1972: 91-106.

Taylor, F., Scientific Management. New York: Harper: 1947

Tullock, G., The Politics of Bureaucracy. Washington: Public Affairs Press, 1965.

Weber, M., The Theory of Social and Economic Organization trans. by A.M. Henderson and Talcott Parsons. Glencoe, Ill. Free Press, 1947.

Williamson, O., The Economics of Discretionary Behavior: Managerial Objectives in a Theory of the Firm. Chicago: Markham, 1967a.

Williamson, O., "Hierarchical Control and Optimum Firm Size", Journal of Political Economy, vol. 75, 1967b: 123-38.

Wilson, R., "Informational Economies of Scale", Bell Journal of Economics, vol. 6, 1975: 184-195.

Wintrobe, R., The Economics of Bureaucracy. Unpublished doctoral dissertation, University of Toronto, 1976.

_____, "On the Nature of the Firm". Paper delivered at the 1977 meetings of the Public Choice Society at New Orleans, March 12, 1977.

Woodward, J., Industrial Organization: Theory and Practice London: Oxford University Press, 1965.

FOOTNOTES

1. Elsewhere, we have shown this more systematically [Wintrobe, 1977]. This argument extends the work of Alchian and Demsetz (1972) on the theory of the firm. Briefly, they showed that firms exist when it is relatively economical to estimate the marginal productivity of factors by observing some index of their input rather than their output; under those conditions if factors were paid according to their output, each factor would have an incentive to shirk, and this shirking imposes external damages on complementary factors. In fact, the shirking discussed by Alchian and Demsetz is simply one form of discretion. It can be shown that other forms of discretion, e.g., in the choice of technical procedures to be followed, in the timing of production ("discipline"), the quality of raw materials, etc., would similarly impose external damages on complementary factors. The only conditions required are (1) that productivity may be increased by the division of labor, and that (2) there are a relatively small number of suppliers of the different sub-processes. These conditions guarantee that the suppliers are mutually interdependent. See Wintrobe (1977).
2. The circumstances which determine the extent to which an organization's technology will be routine or non-routine are discussed in detail in section 4.
3. This is similar to Tullock's "judgemental" system. See Tullock (1965).
4. We assume a "separation theorem", namely that the choice of the optimal level of risk-taking is independent of the division of total risk among the risky "assets" t_h and t_g, i.e., that the choice of b is independent of the choice of q. Alternatively, the analysis can be conducted without this assumption, and is unchanged wherever the results are directly comparable, using $\frac{t_h}{T^*}$ and $\frac{t_g}{T^*}$, instead of b and q as the sub-

ordinates' choice variables. Only the proofs using the first approach will be presented here. Other proofs are available from the author on request.

5. z_i could be made a function of detected violations, but since in any time period, some violations are likely to be rewarded, others punished, it is not obvious whether this function should be increasing or decreasing. The most "reasonable" assumption is that z_i is an increasing function of justified violations, and a decreasing function of unjustified ones. But in that case, z_i should also depend on past violations (those in previous time periods) and the model would become exceedingly complicated. Consequently, we assume that, over the time period T^*, z_i is constant. Interior maxima of (5) and (6) are still guaranteed by the assumption of risk aversion.

6. Applying (5) and (6), we can re-write the first-order conditions in terms of the allocation between states h and f and between states g and f:

$$\frac{h}{f} = \frac{p(1 - z_i)U'_f}{pz_i\, U'_h}$$

$$\frac{g'}{f} = \frac{p(1 - z_i)U'_f}{(1 - p)\, U'_g}$$

We can safely assume that all subordinates are risk averse. The marginal utility of income is therefore everywhere decreasing, and hence since the level of income is higher in states h and g than in state f, U'_h, $U'_g < U'_f$. If the values of q and b are interior maxima, it is required that

$$pz_i\, h > p(1 - z_i)\, f \text{ for } t_h > 0,$$

and

$$(1 - p)\, g > p(1 - z_i)\, f \text{ for } t_g > 0.$$

In the pure authority system discussed in the text, both of these

inequalities are reversed, and $t_h = t_g = 0$.

Second-order conditions require that, for a local maximum,

$$D_1,\ D_4 < 0, \text{ and } \begin{vmatrix} D_1 & D_2 \\ D_3 & D_4 \end{vmatrix} > 0$$

where $D_1 = \dfrac{\partial^2 EU}{\partial q^2} = pz_i h^2 \theta U_h'' + (1-p) g^2 \theta U_g''$

$$D_4 = \frac{\partial^2 EU}{\partial b^2} = pz_i q^2 h^2 U_h'' \theta' + (1-p)(1-q)^2 g^2 U_g'' \theta' + p(1-z_i) f^2 U_f'' \theta'$$

and $D_2 = D_3 = \dfrac{\partial EU}{\partial q \partial b} = 0$ by the separation assumption above. Risk aversion implies that U_h'', U_g'' are both < 0, and hence, these two assumptions guarantee the satisfaction of the second order conditions.

7. See R. Bendix [1974], p. 204.

8. $\dfrac{\partial q}{\partial f} = \dfrac{0 \cdot D_4}{D} = 0$; $\dfrac{\partial b}{\partial f} = \dfrac{D_1}{D}(-p(1-z_i)(fU_f'' \theta - U_f')) < 0$

9. It might be thought that an alternative explanation for the use of relatively low sanctions in many organizations is that, since the maximum sanction available is usually dismissal, this sanction would have to be reserved for the most serious infractions. Less serious infractions would have to be punishable by relatively light sanctions in order to preserve the principle of marginal deterrence. However, Becker and Stigler have shown that the costs imposed on employees by dismissal is not uniform, but can be changed by changing the time-stream of payments embodied in the contract. For example, a decrease in an employees' weekly wage payment, compensated by an increase in the size of his

pension to leave the employee equally well off, would increase the costs of the sanction imposed by dismissal towards the end of the contract. Consequently, if in fact sanctions do tend to be surprisingly low in many organizations, this fact must be explained by the principle of choice and not by the limitations imposed by legal institutions. See Becker and Stigler, (1975).

10. See footnote 1.

11. $\frac{\partial q}{\partial h} = \frac{D_4}{D}(-pz_i U'_h + hU''_h q\theta)$;

$\frac{\partial b}{\partial h} = \frac{D_1}{D}(-pz_i qU'_h + hqU''_h\theta)$

Clearly, $\frac{\partial q}{\partial h}, \frac{\partial b}{\partial h} \gtrless 0$ as

$|pz_i U'_h| \gtrless |pz_i Y_h U''_h q\theta|$, where the term on the left hand side is the substitution effect, and the term on the right hand side is the wealth effect. Moreover, the sign of $\frac{\partial q}{\partial h}$ is always the same as that of $\frac{\partial b}{\partial h}$

$\frac{\partial q}{\partial g} = \frac{D_4}{D}(1 - p)[U'_g + gU''_g(1 - q)\theta]$

$\frac{\partial b}{\partial g} = \frac{D_1}{D}(1 - p)(1 - q)[U'_g + (1 - q)gU''_g\theta]$

Again, the results are ambiguous because of opposing wealth and substitution effects.

12. $\frac{\partial q}{\partial z_i} = \frac{D_4}{D}(-pU'_h hq - pU'_f f) > 0$

$\frac{\partial b}{\partial z_i} = \frac{D_1}{D}(-pU'_f f - pU'_h hq) > 0$

13. $\frac{\partial q}{\partial p} = \frac{D_4}{D}(-z_i U'_h h - U'_g g) > 0$

$\frac{\partial b}{\partial p} = \frac{D_1}{D}(-z_i U'_h hq + (1 - z_i)U'_f f + U'_g g(1 - q)$

< 0 if $|(1 - z_i)U'_f f + U'_g g(1-q)| > |-z_i U'_h hq|$

From the first-order conditions, $pz_i U'_h h = p(1 - z_i)U'_f f$. Multiply both

sides of the inequality by p, and substitute $p(1 - z_i)U'_f f$ for $pz_i U'_h h$. This gives

$$\frac{\partial b}{\partial p} < 0 \text{ if } \left|p(1 - z_i)U'_f f + p(1 - q)U'_g g\right| > \left|-q(p(1 - z_i)U'_f f\right|$$

which is clearly true since $q < 1$.

14. We may also briefly indicate the effects on the level of subordinate discretion of changes in the length of the subordinate's contract, T*, and in k, the number of tasks performed per unit time. Detailed proofs are given in Wintrobe (1976). An increase in T* implies that the dispersion of a subordinate's income per unit time decreases for any given level of discretionary activity T. Consequently, a longer contract allows subordinates to spread the risk of discretionary acitivity over a larger number of risky actions.

A change in k may be shown to have the identical effects on the variance of income per unit time as an increase in T*, and for the same reasons; if we compare two jobs with the same contract length, the job where a larger number of tasks are performed per unit of time permits the risk of a given amount of discretionary time to be diversified over a larger number of "trials". However, diversification is possible only as long as the outcomes of different task performances are not perfectly correlated with each other. An increase in the number of times the same task is repeated in the same way per unit time permits no diversification, and therefore has no effect on risk-taking. k must therefore be interpreted as the number of different tasks performed per unit time.

From the point of view of the effect of contract length on risk-taking, therefore, the effective "length" of a contract is not its length in units of time T*, but the total number of different task performances which the length of the contract allows, kT*. An

Increase in either k or T* increases the effective length of the contract measured this way, and decreases the variance of a subordinate's income per unit time. This decrease in the variance of income per unit time with unchanged expected income exerts a substitution effect in favor of increased risk-taking (discretion), and a wealth effect against it. As in the usual theory of portfolio choice, the net effect will be in favor of an absolute increase in the level of risk-taking if subordinates have decreasing absolute risk-aversion.

15. See the definition of perfect political competition in G.S. Becker, [1976], p. 34.

16. ibid., p. 36, see also G.J. Stigler [1972], p. 97.

17. The term "degree of freedom" is used in this context by Breton [1974].

18. See Breton and Wintrobe [1975]. A much more detailed model of bureaucratic inefficiency in the public sector is currently in preparation.

19. See Baumol [(1959)], Marris [(1964)] and Williamson [(1967b)].

20. Note that this effect of human capital on G' is separate from, and in addition to, the effect of human capital on z_i noted earlier. That is, an increase in human capital clearly increase both the productivity of discretion (G') and the ability of subordinates to justify discretionary action (z_i).

21. Such data are available for the U.S. at least, and used in Blau and Schoenherr [1971] to measure the division of labor.

22. This is the term used by C. Kennedy and A.P. Thirlwall [1972].

23. The basic reference is A. Alchian [1963]. For others, see Rosen [1972], and Kennedy and Thirlwall [1972].

24. Rosen's excellent [1972] paper also discusses the difference between knowledge which is "rested in the firm", to use his terminology, and

that which is rested in the firm's employees.

25. Arrow states that many of the classic learning experiments show that "learning associated with repetition of essentially the same problem is subject to sharply diminishing returns" [1962, p. 155], and uses new investment as a proxy for new stimuli.

26. This process whereby routinization gets built into the operations of a firm on the one hand, and tends to depreciate on the other, also explains why bureaucracies tend to ossify and be so hard to change, why when changes are made they must be sudden and dramatic, and why responsibility is so diffuse within hierarchies. Any instruction for dealing with an exceptional situation tends to become a precedent; faced with a similar situation, the subordinate can insure himself against the consequences of his actions by following this precedent set earlier. When an instruction is initially issued, responsibility clearly lies with the issuer; when a precedent is simply followed, it is not clear where the responsibility lies. The set of precedents which has accumulated inside an organization cannot be dismantled on a piecemeal basis; the expectation must be broken that previous instructions have the status of precedents. This explains why firms and bureaus adapt to even slowly changing circumstances by periodic dramatic re-organizations rather than through slow readjustments.

27. The problem of the "start" of the progress function is discussed in Kennedy and Thirlwall [1972], p. 39, and in Hirsch [1952].

28. Second order conditions for a local maximum require that

$$V_{ff},\ V_{hh},\ V_{pp} < 0,\ \begin{vmatrix} V_{ff} & V_{fh} \\ V_{hf} & V_{hh} \end{vmatrix} > 0, \text{ and}$$

$$D = \begin{vmatrix} V_{ff} & V_{fh} & V_{fp} \\ V_{hf} & V_{hh} & V_{hp} \\ V_{pf} & V_{ph} & V_{pp} \end{vmatrix} < 0$$

By Young's theorem, $V_{fh} = V_{hf}$, $V_{pf} = V_{fp}$ and $V_{ph} = V_{hp}$. We assume that the cross-partial effects are "small" relative to the direct effects, and this assumption may be implemented by the strong condition that V_{fh}, V_{pf} and V_{ph} are all equal to zero. Then the requirement that V_{ff}, V_{hh}, Vpp < 0 guarantees the satisfaction of the other second order conditions as well.

The third second-order condition may be equivalently expressed as $T_f T_h T_p(D)$ where $D = \begin{vmatrix} D_{11} & 0 & 0 \\ 0 & D_{22} & 0 \\ 0 & 0 & D_{33} \end{vmatrix} < 0$

since $T_f T_h T_p = (-)(+)(-) > 0$

and $D_{11} = V_{o_f} - AR_f - \frac{AR_f}{E_f^T} > 0$

$$D_{22} = V_{o_h} - AR_h - h\left[\frac{(AR_{hh} - q_h V_{1_h} - V_{1_{qhh}}) + AR_h - q_h V_1}{E_h^T}\right] < 0$$

$$D_{33} = V_{o_p} - AR_p - p\left[\frac{AR_{pp} - q_{pp} V_1 - q_p V_{1_p} + AR_p - q_p V}{E_p^T}\right]$$

$$- \frac{T_p C_{pp} - C_p T_{pp}}{T_p^2} > 0$$

if the elasticities E_f^T, E_h^T and E_p^T are approximately constant.

29. See, e.g., Blau and Schoenherr [1970].

30. Differences in discretion, expressed as a simple dichotomy between a "routine" and a "non-routine" organizational control system are in fact what lies underneath the diverse typologies which have been proposed by different theorists to explain organizational structure and behavior. Hence Weber's (1947) disctinction between bureaucratic

and charismatic authority, Burns and Stalker's (1961) mechanistic vs. organic systems, Simon's (1960) programmed vs. non-programmed responses, Crozier's (1964) routinize vs. uncertain adaptations, Presthus' (1958) structural vs. unstructured perceptual fields, Taylor's (1947) scientific task determination vs. personal rule-of-thumb, etc. As has been pointed out by Hickson, (1966) in each case, the underlying dichotomy is the same routine - non-routine distinction. A relationship between technology and discretion has been suggested by, among others, Perrow (1967) and Woodward (1965). For examples of sociological measures of discretion, see Hage and Aiken (1969) and references cited there.

31 The measure is, of course, not without its defects. Probably the most important of these is that data limitations undoubtedly preclude taking into account the value of perquisites, prestige, and other non-monetary rewards. The more important such things are, the less successful the measure will be. However, as long as changes in these variables tend to be accompanied by changes in monetary rewards. The measure will be adequate.

A second difficulty is that the dispersion of incomes partly reflects differences in ability (z_i) among employees. If organizations varied in the extent to which subordinates at the same level differ in their abilities, then differences in income dispersion might be due to this factor rather than to differences in the average levels of discretion. However, the measure will still be unbiased, as long as differences in z_i among employees do not vary systematically with changes in the exogenous variables which we use to explain dispersion. There is no obvious reason why this should be so. The effect of this consideration is therefore simply that some of the differences in

dispersion among organizations will be left unexplained by the present model. Our purpose, of course, is not to explain dispersion, but discretion, and as long as changes in discretion do give rise to differences in income dispersion, the measure will suit our purpose.

32. $$\frac{\partial \hat{f}}{\partial y_1} = \frac{D_{22}D_{33}}{D} H_{1_{T_{Y_1}}} = (\frac{(-)(+)}{(-)}(+) > 0$$

$$\frac{\partial \hat{h}}{\partial y_1} = \frac{D_{11}\ D_{33}}{D} H_{IT_{y1}} = \frac{(+)(+)}{(-)}(+) < 0$$

$$\frac{\partial \hat{p}}{\partial y_1} = \frac{D_{11}D_{22}}{D} H_{IT_{y1}} = \frac{(+)(-)}{(-)}(+) > 0$$

where y_1 is an exogenous variable which increases H_{I_T}

33. If y_2 is an exogenous variable which increases the marginal gain from productive discretion, G', the optimal values $\hat{f}$, $\hat{h}$, and $\hat{p}$ change as follows:

$$\frac{\partial \hat{f}}{\partial y_2} = \frac{D_{22}D_{33}}{D}(-qG'y_2) < 0; \quad \frac{\partial \hat{h}}{\partial y_2} = \frac{D_{11}D_{33}}{D}(-q - \frac{hq_h}{E_h^T})\ G'y_2 > 0$$

since both $q_h > 0$ and $E_h^T > 0$ in the relevant region, as shown in the text.

$$\frac{\partial \hat{p}}{\partial y_2} = \frac{D_{11}D_{22}}{}(-q - \frac{pq_p}{E_p^T})G'y_2$$

The term in the brackets may be written as $-q(1 + \frac{E_p^q}{E_p^T})$

For any subordinate $q_p > 0$, $T_p < 0$ as shown previously. Hence $E_p^q > 0$, $E_p^T < 0$, and $\frac{E_p^q}{E_p^T} < 0$. If $\left|\frac{E_p^q}{E_p^T}\right| > 1$ the term in the brackets is negative and then $\frac{\partial \hat{p}}{\partial y_2} > 0$ unambiguously.

$\left|\frac{E_p^q}{E_p^T}\right| > 1$ if $|q_p| > |T_p|$ and $\left|\frac{p}{q}\right| > \left|\frac{p}{T}\right|$; $|q_p| > |T_p|$, since $q_p > 0$, $T_p < 0$ and $\frac{\partial(qT)}{\partial p} \equiv \frac{\partial t_h}{\partial p} > 0$; $\left|\frac{p}{q}\right| > \frac{p}{T}$ since $q < 1$, $T > 1$

Hence $\left|E_p^q\right| > \left|E_p^T\right|$, and $\frac{\partial \hat{p}}{\partial y_2} > 0$.

34. E_p^T will be small relative to E_f^T for two reasons. First, if employees are risk-averse, $E_p^{tg} < E_f^{tg}$, as shown in Becker's (1968) paper on crime and law enforcement. (Becker showed risk-aversion implied that $E_p^O < E_f^O$, where O is the number of criminal offenses, and f the criminal sanction). Second $E_p^T = E_p^t h + E_p^t g$, and $E_p^t h > 0$, $E_p^t g < 0$, while $E_f^T = E_f^t h + E_f^t g$ and $E_f^t h$, $E_f^t g$ both < 0. That is, the changes in t_h and t_g due to the change in p are each smaller than the changes due to the change in f, and offset rather than reinforce each other as to the changes in f. Hence a fortiori, $E_p^T < E_f^T$.

35. See Meyer (1968) and Woodward (1965).

36. See Meyer (1968) and Woodward (1965).

37. An exogenous increase in H_2' will change $\hat{f}$, $\hat{h}$ and $\hat{p}$ as follows:

$$\frac{\partial \hat{f}}{\partial y_5} = \frac{D_{22}D_{33}}{D}(1 - q)\, H_{2_{y_5}}' > 0$$

$$\frac{\partial \hat{h}}{\partial y_5} = \frac{D_{11}D_{33}}{D}\left[(1 - q) - \frac{hq_h}{E_h^T}\right] H_{2_{y_5}}' > 0 \text{ if } (1 - q) - \frac{hq_h}{E_h^T} < 0$$

Since $(1 - q) < 1$, $(1 - q) - \frac{hq_h}{E_h^T} < 0$ if $\frac{hq_h}{E_h^T} > 1$.

We prove $\frac{hq_h}{E_h^T} > 1$ as follows.

Since $\frac{\partial(1 - q)T}{\partial h} \equiv \frac{\partial t_q}{\partial h} < 0$, and since $T_h > 0$ and therefore $(1 - q)_h < 0$.

Hence $\frac{\partial(1 - q)T}{\partial h} < 0$ implies $q_h > T_h$.

If $q_h > T_h$, $\frac{hq_h}{E_h^T} \equiv \frac{q_h}{T_h} T > 1$ and therefore $\frac{\partial \hat{h}}{\partial y_5} > 0$.

Finally, $\frac{\partial \hat{p}}{\partial y_5} = \frac{D_{11}D_{22}}{D}[(1 - q) - \frac{pq_p}{E_p^T}]H_{2y_5}' > 0$ since $E_p^T < 0$, $q_p > 0$.

38. An exogenous increase in C_p would not change the optimal values of f or h, since

$$\frac{\partial \hat{f}}{\partial Y_3} = \frac{D_{22}D_{33}}{D}(0) = 0, \text{ and } \frac{\partial \hat{h}}{\partial y_3} = \frac{D_{11}D_{33}}{D}(0) = 0$$

The effect on the optimal value of p is given by

$$\frac{\partial \hat{p}}{\partial y_3} = \frac{D_{11}D_{22}}{D} \frac{1}{T_p} C_{py_3} < 0 \text{ since } T_p < 0,$$

The effects of an exogenous change in C_T are given by:

$$\frac{\partial \hat{f}}{\partial y_4} = \frac{D_{22}D_{33}}{D} C_{Ty4} > 0; \quad \frac{\partial \hat{h}}{\partial y_4} = \frac{D_{11}D_{33}}{D} C_{Ty4} < 0; \quad \frac{\partial \hat{p}}{\partial y_4} = \frac{D_{11}D_{22}}{D} C_{Ty4} > 0$$

if $C_{Ty4} > 0$.

39. While both these effects increase the costs of controlling discretion, they have opposing effects on C_p, and therefore on the efficiency of reducing T by increasing p. The evidence consistently shows a decline in the ratio of administrative personnel to direct labour as size increases, but this evidence is equally consistent with the dominance of either sampling economies (implying an increase in p) or control loss (implying a decrease). For samples of the evidence, see e.g. Blau and Schoenherr (1970)(who favor the former interpretation) and Pondy (1969) (who favors

the latter).

40. The most recent - and the most precise - statement of the classical theory of firm size - that by Williamson [1969] bases the limit to firm size on the control loss and administrative overload factors alone, and implicitly assumes that there are no scale economies in information collection.

41. The average costs of obtaining information on any routine characteristic are proportional to $\frac{M}{N}$, where $\underline{n}$ is the size of the sample, and N the number of subordinates who share the characteristic, i.e. perform the same routine operation. Average information costs do fall continuously as N increases. But clearly the size of this reduction in average costs diminishes with larger and larger N.

42. The marginal productivity of discretion diminishes as the firm ages, via the process described above. Hence, with age, the loss in productivity due to the reduction in discretion as size increases will be smaller, and hence the firm's optimum size increases with age. This explains why competitive firms tend to grow over time. Some theorists have further argued, on similar grounds, that only a firm's rate of growth, and not its absolute size, is limited by difficulties of coordination and control. (e.g. Penrose (1959), Rubin (1973)).

There is, however, a limit to how large a firm can grow, since our analysis shows that the firm can only continue to grow by becoming more and more routine, and this is the point which is missed in these arguments. As long as there is a limit to how routine the firm can become without encountering increasing average costs, there is an upper limit to the size as well as the rate of growth of the firm.