Budgets as Dynamic Gatekeepers

Harold Pollack • Richard Zeckhauser
Yale University, Institution for Social & Policy Studies, 89 Trumbull Street,
PO Box 208207, New Haven, Connecticut 06520-8207
Kennedy School of Government, Harvard University, 79 John F. Kennedy Street, Cambridge, Massachusetts 02138

Most large organizations allocate resources by means of fixed budgets: each subunit is normally entitled to spend a defined amount over a fixed period, usually one year. Fixed budgets create clear incentives for subunits to control costs. Yet such arrangements create major incentives for dynamic inefficiency, for example by encouraging subunits to exhaust their budgets toward the end of the fiscal year.

This paper develops a dynamic optimization model to examine the incentives fostered by budget systems. It invokes the metaphor of physicians involved in a health care delivery system to examine incentives created by decentralized "gatekeeping" as a mechanism to control medical costs. The paper also discusses some methods to reduce the incentives for dynamic inefficiency that fixed budgets create.

(*Agency Theory; Budgeting; Dynamic Programming; Gatekeeping; Resource Allocation; Rationing*)

1. Introduction

In most large organizations, central authorities lack critical information and expertise that is available to decision-makers lower down in the hierarchy. Hence, they decentralize decisions. A principal instrument for such decentralization is the budget: a central planner assesses the expected productivity of marginal expenditures in different subunits, and then sets a fixed, per-period budget for each subunit designed to equate these productivities. The choice of specific projects or endeavors to fund is delegated to decision-makers in subunits, individuals we refer to as gatekeepers.

In making such decisions, gatekeepers respond not only to the overall mission of the organization, but also to a number of perverse budgetary incentives to reduce efficiency. This paper explores how rational gatekeepers respond to these incentives. It also examines how budgetary strategies could be altered to reduce the incentive and risk-sharing difficulties common to decentralized settings.

We invoke the metaphor of physicians who work in a health care delivery system in which the central authority seeks to control overall expenditures by placing a cap on the amount that can be spent by each physician. Given the rapid growth of medical expenditures in most developed nations, and the central policy concern with health care cost containment in the United States, this metaphor seems timely. The analysis could also easily be applied to a foundation handing out awards, a social service agency taking on new clients, a corporate unit hiring engineers, or a professional school hiring junior faculty in a rolling search. In each case, the subunit is authorized by higher authorities to spend a given amount within a given period, usually a year.

For illustration, consider a specific cost-containment scheme that was proposed by Britain's National Health Service (NHS) during the latter Thatcher years. It would have established an account for each general practitioner (GP) that would be charged every time he referred a patient to the hospital, the amount depending on the complexity of the referral.[1] The NHS plan asked physicians to lower health costs by being medical gate-

[1] To facilitate exposition, we shall assume that gatekeepers are men, and that the central planners or higher authorities are women.

0025-1909/96/4205/0642$01.25

keepers, to weed out low priority uses of NHS resources.[2] Health care organizations in the United States have instituted similar arrangements in an effort to control rising costs. For example, Medicaid gatekeeper initiatives in California and New Jersey have been estimated to reduce visits for specialty care by almost fifty percent.[3]

Such gatekeeping arrangements, though best known in medical organizations, are widespread in other settings. Public educational services, university admissions in the nonprofit arena, and private-sector bank loans are all implicitly or explicitly rationed through gatekeepers' decisions. The arrangements may be informal: Lipsky (1987) shows how police officers and social workers effectively ration public services.[4]

In a broader context, the collapse of Communism in Eastern Europe illustrates the need for efficient rationing below the level of any central authority. As Friedrich Hayek (1945) noted long ago, command economies were bound to be inefficient because central planners could not possibly collect and analyze all the information reflected in local market prices. Contemporary management texts emphasize that Hayek's insight applies with equal force within any large organization: leading decision-makers inevitably lack critical information and expertise available to lower-level agents.[5]

Yet decentralization creates distortion of its own. Except in unusual cases where preferences among subunits completely coincide, decentralization produces familiar principal-agent incentive problems.[6] While those who staff organizations value the mission of the employer, there is not a complete coincidence of motives or material interest. Organizational subunits and individual professionals are more concerned with local matters than with wider goals. Most doctors care far more about their own patients than they do about the well-being of the overall health system. The same "hometown concern" is found among chairs of university academic departments, and among administrators of state aid programs.

Moreover, efficiency dictates that an organization's resources be channelled to the most productive subunits. Yet productivity is notoriously difficult to measure, especially within the public and nonprofit sectors where there is rarely an accepted measure of output and where many inputs are not priced at all or prices do not reflect true marginal costs. Central planners must often estimate output and cost figures. These estimates are often derived from information supplied by the subunits themselves, which have obvious incentives to depart from candor.

Further distortions arise as uncertainties are resolved over time. Although we usually think about resource allocation as a static process covering a short period, in fact budgets are generally determined for use over a longer fixed period, which we will call the *fiscal year*. Such a budget strategy offers little difficulty if subunits are large enough so that each faces a small proportional variance in its optimal expenditure due to unexpected contingencies. This method works less well when subunits face large or highly correlated risks, or when the fiscal year is too short to smooth out variability due to repeated chance occurrences. For example, it is undesirable to set a rigid budget for the highway department of a northern city where unpredictable winter weather affects the amount of snow that must be removed and the damage done to roads.

However careful the optimization at the outset, when local variability is great, different subunits will face

[2] Kumar et al. (1993) provide simulations highlighting the role of decentralization and associated externalities in a static context. Malone (1987) provides a useful overview of coordination issues.

[3] See, e.g., *American Medical News* (1994).

[4] A referee aptly notes that much U.S. entitlement spending, such as Medicaid and Medicare, is not constrained by fiscal year limits. The gatekeepers model would nonetheless apply if annual caps were placed, say on hospital expenditures. Similar caps were part of the Clinton administration's proposed health care reform, primarily as a source of funding to subsidize new beneficiaries. Congressional Republicans show interest in capitated payments for providers as a tool to control entitlement spending. Given present U.S. programs, the gatekeepers model readily applies to efforts that give out loans or grants, say for economic development or NSF research.

[5] Milgrom and Roberts (1992) provide management case studies informed by information economics.

[6] Examples in which decentralization does not produce distortion include (perhaps) sports teams, and firms in which subunits can be charged the marginal cost of scarce resources, such as when profit centers bid for the services of the central marketing department.

dramatically different shadow prices for marginal expenditures once these uncertainties have been resolved. This makes it efficient to reallocate funds to the most productive subunits during the fiscal year. Yet to do so is especially difficult when subunits derive important parochial benefits from their expenditures, if only a reduction in the political heat from clients and constituents. Except in rare circumstances, we must expect that each unit will virtually exhaust its available resources during every fiscal year, often proclaiming that it should have received more.

Principal-agent theorists have developed mechanisms that seek to minimize such distortions by using side-payments to induce each subunit to reveal its true valuation of scarce resources. Others have focused on tournaments and yardstick competition. Such relative performance mechanisms reduce the informational burdens facing central planners in settings where absolute productivities are difficult to measure.[7]

In this paper, however, we focus on simple budgetary schemes, because most organizations have found budgets to be the appropriate principal mechanism for balancing centralized control and decentralized allocation. In deploying their budgets, most decentralized organizations, particularly in government and the nonprofit sector, distribute scarce dollars, personnel, and support services based upon ex ante estimates of each subunit's productivity. Such estimates may be based upon studies by a central budget office or legislative committee, but past expenditures or explicit formulas such as dollars per client served usually play a major role. While many organizations provide a separate channel to provide for emergencies or other unforeseen events, fixed *ex ante* budgets are the norm in running government agencies, many non-profits, and even for-profit organizations.

Fixed budgets predominate because they perform several important organizational tasks. Budgets communicate priorities through the chain of command, establishing criteria by which each subunit will be held accountable. Wildavsky (1974) emphasizes that budgets are also political documents; they reflect compromises between a government agency and the legislature, as well as internal compromises among conflicting constituents and goals. Budgets provide a workable, if imperfect, strategy to balance these tasks and concerns.

For these reasons, scholars have closely analyzed the role of fixed budgets in many settings. Niskanen (1971), employing a principal-agent framework, argued that public expenditures tend to be excessive because budget-maximizing bureaucrats exploit their superior knowledge—i.e., extract informational rents—in their dealings with superiors and overseers.

During the 1980s, many scholars extended and scrutinized Niskanen's work, often focusing on the interplay between bureaucratic accountability and administrative structure. Within one model, rational legislators are mindful that agencies may exploit their informational advantages to escape effective oversight. Legislators also know that it is costly to observe what agencies actually do. One way to control agencies at low monitoring cost is to structure decision-making so that relevant interest groups have the appropriate incentives to act as monitors, say as plaintiffs in real or threatened legal proceedings. As McCubbins and Schwartz (1984) describe this "fire alarm" paradigm, legislators can intervene when constituencies report that executive agencies have departed from legislative directives. Organized interest groups derive influence from the combination of their access to information that is helpful to elected politicians, and their ability to "pull an alarm."

McCubbins et al. (1989) examine ways that legislators establish ex ante procedural constraints to forestall subsequent political gaming by bureaucrats. McCubbins, Noll and Weingast do not explicitly discuss gatekeeping or other budgetary regimes. Yet in light of their model such instruments may be understood as procedural efforts to prevent agencies and organizational subunits from departing from the priorities of their legislative overseers.

Other analyses explore the use of ex post auditing of subunits' expenditures.[8] For example, central planners

[7] On incentives in teams, see Marschak and Radner (1972), Holmstrom (1982), and Pratt and Zeckhauser (1987). On the use of relative performance measures, see Weimer (1992) and Milgrom and Roberts (1992).

[8] For example Banks (1989), Bendor et al. (1987), Banks and Weingast (1992). For an economic survey, see Besanko and Sappington (1987).

may examine a subunit's records if its spending over the fiscal year greatly exceeds prior estimates. In contrast with gatekeeping, ex post audits do not impose prior restrictions on subunits, and therefore do not lead to the characteristic inefficiencies associated with fixed budgets.

However, the threat of audit loses credibility if monitoring costs are high or if ex post penalties are difficult to impose. In such cases, a rational central planner will investigate only the most egregious cases, those for which there is a very high probability that a subunit has departed from the center's wishes. This renders auditing less effective than gatekeeping or other budgetary schemes; an auditing regime allows subunits to shade at the margin, that is to pursue slightly inappropriate expenditures unlikely to be audited.

Important empirical literature on public budgeting and hierarchy helps to frame this paper. Such historic accounts as Fisher (1975) and Wildavsky (1974) describe the complex legal and administrative history of federal budgeting, including efforts by the President and Congress to manage contingency funds, reprogramming of expenditures, and other strategies designed to address unforeseen events. These studies highlight the importance of predictability and clear accountability to Congress as fundamental constraints on novel budget approaches. More recently, Meyers (1988) and Kearns (1993, 1994) have considered the experiences of state governments that employ biennial budgets. These analyses are discussed below, in light of the dynamic analysis of §4.

1.1. Outline of the Paper

The next section presents the basic budgeting problem in a decentralized static setting, examining cases where the central planner can and cannot directly observe subunits' productivities.

Section 3 examines the intertemporal problems faced by gatekeepers in a decentralized setting, developing a simple dynamic programming framework to assess distortions in gatekeepers' decisions due to variability in the demand for resources over time. Section 4 provides some illustrative graphs and calculations for specific examples.

Sections 5 and 6 then discuss strategies to ameliorate the dynamic inefficiencies imposed by decentralized gatekeeping, and consider some practical obstacles to efficient budgeting.

2. Decentralized Budgeting in a One-period Model

Rather than focus on information asymmetries between a public agency or firm and its nominal superiors, this paper examines a basic agency problem that occurs *within* many decentralized organizations. An agency has a fixed budget—say $B—to spend over the fiscal year. The head of that agency wishes to channel these funds to the most valuable uses. Yet each of her subordinate gatekeepers faces obvious incentives to spend too much. What should she do?

Before proceeding to a more general analysis, it is useful to consider the simplest case in which there are no incentive problems or uncertainties that unfold over time. Given these assumptions, optimal budget policies for decentralized gatekeeping are easy to describe. The central planner should simply equate the expected marginal contribution of each dollar at every organizational subunit, subject to the organization's total spending constraint.

More formally, suppose that there are n subunits receiving a budget allocation $\mathbf{b} = (b_1, \ldots, b_n)$. From the perspective of the center, each subunit i has some known production function G_i which depends both upon its budget b_i and the prevailing state of the world. We denote this by the state vector $\mathbf{s} = (s_1, \ldots, s_m)$, distributed according to some density function $p(\mathbf{s})$.[9]

If decisions must be made before $\mathbf{s}$ is observed, then central planners will pick $\mathbf{b}$ to maximize

$$\sum_{i=1}^{N} \int p(\mathbf{s}) G_i(b_i, \mathbf{s}) d\mathbf{s}, \qquad (1)$$

subject to the overall budget constraint

$$\sum_{i=1}^{N} b_i \leq B. \qquad (2)$$

[9] For some concerns—for example nation-wide inflation—the state of the world might affect all subunits equivalently. In contrast, recession may cut a city's highway costs, but will raise costs for the welfare department.

To assure interior solutions, we suppose that the production function G_i is smooth and strictly concave, with

$$\frac{\partial G_i(b_i, s)}{\partial b_i} > 0, \quad \frac{\partial^2 G_i}{\partial b_i^2} < 0, \quad \lim_{b_i \to 0} \frac{\partial G_i(b_i, s)}{\partial b_i} = \infty. \qquad (3)$$

Given these assumptions, planners will equate the expected marginal product of each subunit, or

$$E_s\left[\frac{\partial G_i}{\partial b_i}\right] = \int p(s)\, \frac{\partial G_i(b_i, s)}{\partial b_i}\, ds = \lambda. \qquad (4)$$

Here λ is a Lagrange multiplier that represents the expected marginal utility of an additional dollar in the overall budget.

In the limiting case, where **s** is common knowledge, the budget can be allocated so that the marginal output gained for the last dollar spent by each subunit is the same. Over time, the central authority can revise budgets as information becomes available. Alternatively, it could enforce criteria by which subunits make expenditures.

Our analysis considers the more realistic situation in which the productivity of expenditures is at least partly private information held by each subunit. (The existence of such private information is the major justification for decentralization to begin with.) For instance, the administrator of a neighborhood clinic may know that demand has been slack; hence a large budget cut would have a rather small impact on his patients. Yet he is unlikely to share this insight with his superiors. Such information asymmetries thrust the central planner into a second-best situation: she must seek to equate subunits' productivity at the margin.

3. Uncertainties Unfolding Over Time

Many important uncertainties are resolved over the course of the fiscal year. This presents an important dynamic problem that gatekeepers and their superiors commonly confront as information unfolds.

Under the usual scheme, a gatekeeper is given authority to expend a certain amount within a budgetary cycle. Unspent funds are returned to the central office at the end of the cycle, providing little or no additional utility to the gatekeeper. Because of random fluctuations, the gatekeeper is uncertain about future demand. Near the beginning of the fiscal year, he will therefore be reluctant to expend his funds at an appropriate rate because he has no emergency reserve, and fears that he will be unable to serve especially worthy clients or projects toward the end of the budget cycle. Should no such clients be forthcoming, the gatekeeper will then have an incentive to serve marginal candidates he would have rejected earlier on.

Such incentives for wasteful "spend-downs" are aggravated by the complicated signalling game officials play with superiors and other audiences. Underspending sends an unfortunate message about one's budgetary requirements for future periods. Ironically, it may also be taken as a sign of incompetence, a sign that projects are behind schedule or that needed work has not been done. From a signalling perspective, often the best strategy is to create over-runs that bolster claims that one has a high shadow price for expenditures. In this spirit, one official, dismayed that her department had failed to use its full budget authority during the fiscal year, admonished her staff:

> While I know that it is difficult to spend every last penny and some lapses are always going to occur, the magnitude of these lapses makes me extremely uncomfortable for three reasons:
>
> (1) . . . some astute budget analyst is going to notice a pattern and reduce our appropriations, figuring that if we don't spend it we must not need it.
>
> (2) significant underspending makes it difficult to justify any increases in these appropriations . . .
>
> (3) I am concerned that clients needing services are not getting them.[10]

3.1. A Dynamic Programming Model

In this section, we leave aside issues of political gaming; that is, we assume that the gatekeeper spends all of this year's budget, and that the pattern of spending has no effect on next year's budget. Given this restriction, how would a gatekeeper, facing the dynamic incentives inherent in finite-period budgeting, choose to allocate the resources at his disposal? The model below captures the most important features of the gatekeeper's decisions.

[10] *Washington Monthly*, 1987.

Suppose that a doctor is given a fixed budget of $\$b$ for the fiscal year. The fiscal year consists of T periods, $t_1 \ldots t_T$. Within every period t, one patient arrives who can derive benefit z_t from the doctor's services. The doctor observes z_t, and then decides whether or not to provide a referral for further treatment. There are no call-backs. If the doctor decides not to provide the referral at t, he cannot subsequently revisit this decision.[11]

To simplify the analysis, we assume that referral costs are normalized to be a discrete unit cost of \$1 for all patients. One might therefore think of z_t as the cost-effectiveness of the treatment ultimately provided. We also assume that the collection $\{z_t\}$ consists of independent, identically distributed random variables, each with some cumulative distribution $F(z)$ admitting positive density $f(z)$. Every patient gets nonnegative benefit from the referral; so $z \geq 0$.

For the moment, we ignore discounting, which is not likely to be a major consideration over a twelve-month fiscal year. In multi-year contexts, however, discounting may be an issue. We consider multi-year budgets as a possible reform; we return to discounting at the end of this section.

Under these assumptions, one can formulate the doctor's decision as a dynamic programming problem. Denote his remaining resources as x and the time as the variable t. (The budget gives the level of initial resources; thus $x(0) = b$.)

The doctor will provide a referral whenever a patient will receive benefit greater than some cutoff $C(x, t)$, which depends on his remaining resources x, and the remaining time. $C(x, t)$ may be interpreted as the shadow price of expending one unit of the doctor's budget now, rather than waiting to expend that unit optimally on later patients.[12]

We also assume that the doctor's payoff is proportional to the benefit z_t provided to the patient. One can interpret this to mean that the doctor is seeking to maximize his patients' (though not necessarily taxpayers' or other patients') welfare. This assumption is intended to highlight that many budgetary problems stem from the legitimate, if parochial interests of divergent actors rather than from the simple desire for personal gain. Another interpretation would be that the doctor receives some explicit or implicit proportional payment for his decisions.

If the doctor treats all patients whose condition lies above the cutoff $C(x, t)$, then with probability $1 - F(C)$ the doctor will treat a patient at t and will provide expected benefit $E[Z \mid Z > C]$. Once the doctor provides this referral, he can only refer $x - 1$ patients in the remaining time. On the other hand, with probability $F(C)$ the doctor provides no referral, and can therefore still refer x patients in the remaining periods.

If the doctor acts to optimize his expected benefit $V(x, t)$, V satisfies the Bellman recursion:

$$\begin{aligned} V(x, t) = {} & [1 - F(C)][E[Z \mid Z > C] \\ & + [1 - F(C)]V(x - 1, t + 1) \\ & + F(C)V(x, t + 1). \end{aligned} \tag{5}$$

Given (5), how can we find the optimal threshold? At an interior solution, we can differentiate with respect to C, which yields the optimal shadow price relation

$$C(x, t) = V(x, t + 1) - V(x - 1, t + 1). \tag{6}$$

This condition is fully intuitive; the doctor will provide a referral as long as the expected benefit to the patient exceeds the option value of being able to refer someone else later.

Clearly, we expect the threshold to be highest when resources are limited and much time remains. If, for instance, the doctor can only refer one patient during the entire period, he will set a very high initial threshold, waiting for a patient who is on the extreme upper tail of the distribution. At the end of the budget period, because there are no budgetary carry-overs, the doctor has the incentive to refer everyone he can. So the threshold drops to zero near the end of the budgetary period, yielding the boundary conditions

$$V(0, t) = 0, \quad V(x, T - a) = aE[Z], \quad x \geq T - a. \tag{7}$$

[11] For an analysis of formally similar dynamic problems in other contexts, see Mondschein (1993).

[12] This presupposes basic ethical and professional norms. An anonymous reader recounts that Quebec surgeons are highly paid for a fixed number of patients, and are then reimbursed at a much lower rate for subsequent work. The problems this creates are illustrated by the quip "If you need surgery in November, go to the Bahamas."

We can also apply this analysis to the case of Poisson arrivals in continuous time, often an appropriate framework for arrival problems. Suppose that over the interval $(t, t + dt)$ a patient arrives at the office with probability $\mu^* dt$. The probability of more than one arrival over this interval is assumed negligible.

After manipulation, the Bellman recursion becomes:

$$\frac{\partial V(x, t)}{\partial t} = -\mu[1 - F(C)](E[Z \mid Z > C] - C), \quad (8)$$

$$C(x, t) = V(x, t) - V(x - 1, t), \quad \text{and} \quad (9)$$

$$V(0, t) = 0, \quad C(x, T) = 0. \quad (10)$$

Since $V(0, t) = 0$, Equation (9) implies that $C(1, t) = V(1, t)$.

It is rarely possible to find closed-form solutions in either the Poisson or the discrete case. However, numerical solutions are easily computed. The Poisson case admits closed-form solution in one important special case. If the benefit distribution is exponential—$F(z) = 1 - \exp(-\lambda z)$—we have

$$V(x, t) = \frac{1}{\lambda} \log\left[\sum_{m=0}^{x} \frac{\mu^m (T - t)^m}{m!} \right]. \quad (11)$$

Notice that when x is relatively large, $V(x, t)$ is approximately $\mu(T - t)/\lambda$. In other words, when the budget constraint is unlikely to bind, the expected benefit is equal to the expected number of arriving patients—$\mu(T - t)$—multiplied by the per-patient expected benefit $(1/\lambda)$. This limiting condition holds for more general class of distributions, as is apparent by applying a cutoff of close to zero in the differential equation (8).

"Fat-tailed" distributions for the value of referrals yield higher expected values to the gatekeeper than more tightly packed distributions with the same mean. More formally, increasing the variance of the benefit through a mean-preserving spread will increase the gatekeeper's expected payoff; it increases the option value in selecting who will be referred. This is formalized below:

PROPOSITION 1. *Let X and Y be two random variables with the same mean and the additional property that Y is obtained from X by a mean-preserving spread, $X + \epsilon$, where ϵ has conditional mean zero given X. This is equivalent to the statement that X exhibits second-order stochastic dominance over Y.*

Then if V_X and V_Y are the value functions that correspond to the two distributions, $V_Y(b, t) \geq V_X(b, t)$ for all budgets b and times $t \leq T$.

PROOF. See Appendix.

3.2. Discounting and Interest Payments

The above dynamic optimization can be modified to include the possibility of discounting and interest payments at rate r on as-yet unspent funds. For ease of exposition, we consider only the case of one arrival per fixed-length period. The case of Poisson arrivals can be treated analogously.

We assume that the gatekeeper has a per-period discount rate ρ, defined so that z units of benefit in period $t + 1$ is equally valued as only $z/(1 + \rho)$ units at time t. For completeness, we also assume that gatekeepers receive interest payments on as-yet unspent funds. If the doctor has x available referrals at period t, he will have $x(1 + r)$ available at period $t + 1$ if the patient who arrives at period t is not referred. If that patient is referred, the gatekeeper has a remaining budget of $[1 + r](x - 1)$. The effective "cost" of a referral is the chance to make $(1 + r)$ referrals in the next period. We ignore the integer constraint, which makes sense if x is large or if it is possible to provide fractional treatments with proportionate results.

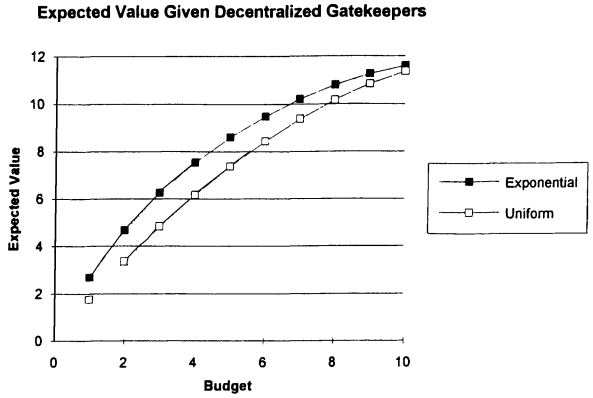
As before, the gatekeeper wishes to maximize his expected benefit, $V(x, t)$. With discounting and interest, V satisfies the Bellman recursion:

$$V(x, t) = [1 - F(C)][E[Z \mid Z > C] + \frac{([1 - F(C)]V([1 + r](x - 1), t + 1) + F(C)V([1 + r]x, t + 1))}{1 + \rho}. \quad (5a)$$

As before, we differentiate with respect to C to derive the shadow price relation

$$C(x, t) = \frac{V([1 + r]x, t + 1) - V([1 + r](x - 1), t + 1)]}{1 + \rho}. \quad (6a)$$

Figure 1A Expected Value Given Decentralized Gatekeepers



Again, the doctor will provide a referral as long as the discounted expected benefit to the patient exceeds the option value of a later referral. At any time before the end, and with any level of resources, it is not surprising that a high discount rate leads the gatekeeper to be less selective in his referrals. In practice, budgetary periods in the United States and other developed countries are rarely long enough to make discounting a serious concern. From the perspective of the social planner, a plausible discount rate is the real return on U.S. government securities. Surprisingly, Ibbotson and Sinquefield (1989) document that real interest rates in the U.S. over the past sixty years have averaged roughly 1% per anum.

4. Values and Thresholds for Representative Examples

We shall now explore a range of examples, looking at the nature of optimal strategies and the values they yield. All the examples, unless otherwise noted, deal with the discrete case with twelve periods (corresponding to months).

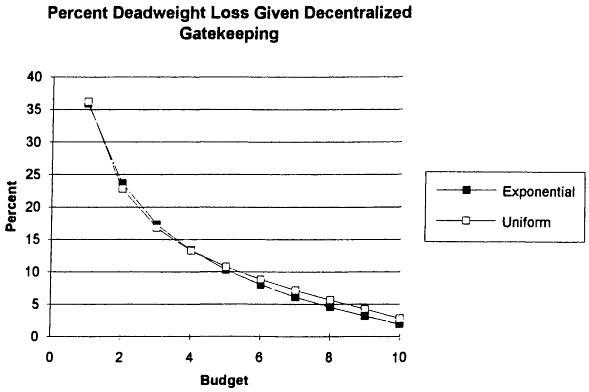
4.1. Expected Values

The top of Figure 1 shows the gatekeeper's expected values for the exponential distribution with mean one, and for the uniform distribution on [0, 2], which has the same mean. The figure shows the expected values for various budgets, assuming that each gatekeeper maximizes his own expected value over the twelve periods. The uniform distribution exhibits second-order dominance over the exponential. Hence, by Proposition 1 the exponential yields a higher expected payoff to the gatekeeper.

As shown in the bottom of Figure 1, such decisions produce substantial deadweight loss relative to what would be achieved with the same overall budget if the central planner could set a cutoff that applies to all subunits over all periods, as she would with fully efficient decentralization. (With the social optimum, the budget is an expected expenditure, with some subunits spending more, some less.) In percentage terms, the deadweight loss is largest when the overall budget is small—that is, when the gatekeepers must be especially selective in the provision of care. Deadweight loss quickly drops as the budget constraint is loosened. In the limit, as the budget constraint is removed, there is no deadweight loss from decentralization since the gatekeeper can refer every patient.

One can readily calculate sample paths of the referral threshold $C(x, t)$ for representative benefit distributions. Along any given path, as time goes by and no patients are treated, the referral threshold will drift down. Whenever a patient is treated, the threshold jumps up. Near the end of the fiscal year, the doctor may have to set a threshold well below social cost to make sure to spend his full budget. In the last period, if he has any remaining resources, he sets a threshold of zero. Thus, when the time horizon is short, physicians depart markedly from the social optimum.

Figure 1B Percent Deadweight Loss Given Decentralized Gatekeeping



The dynamic programming model developed above has several implications for decentralized decisions. If decision-makers follow such individually rational strategies, they will fail to produce the social optimum for at least two reasons:

- Gatekeepers will tend to hoard their budgets early on, for fear of running out later. So in early periods, care is not dispensed even when the benefit to patients exceeds the social cost.
- Individually-optimal strategies produce socially wasteful "spend-downs" toward the end of each budget cycle.

4.2. Simulated Results for the Exponential Distribution

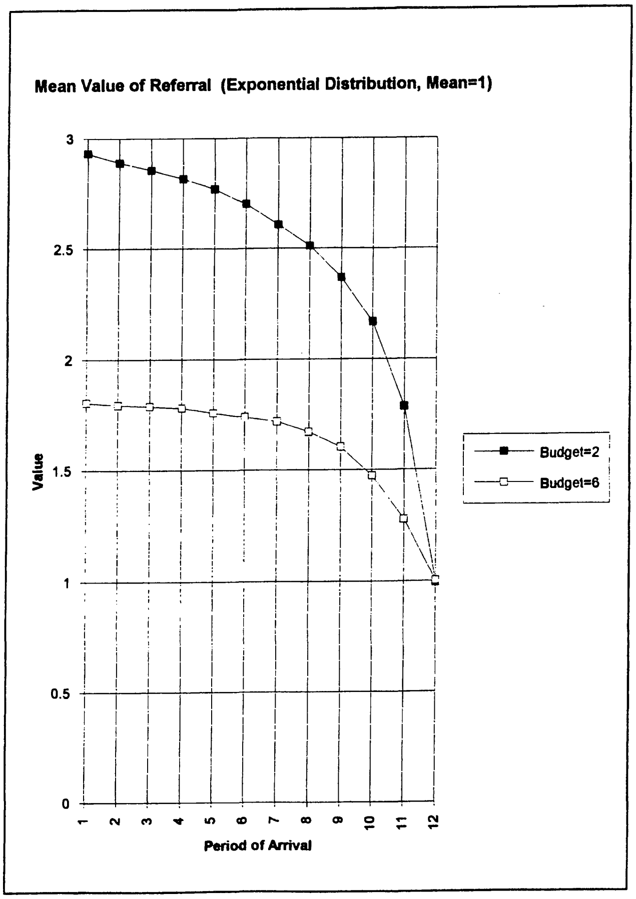
To give a feel for qualitative results, we employed a Monte Carlo routine to simulate 100,000 trials with the exponential distribution. (An earlier simulation with 3,000 trials gave surprisingly choppy results in some instances.) Figure 2 shows the mean value of a referral for total budgets of 2 and 6. The expected benefit reaped by a referral rapidly declines as the end of the fiscal year approaches. (Simulations of the uniform case yield qualitatively equivalent results.)

Figure 3 examines these issues in greater detail by displaying the referral probability for patients when the gatekeeper can make a total of two referrals over the twelve periods, with one patient arriving per period. The graph was derived via Monte Carlo simulations, using 100,000 trials to assure convergence. The case in which the gatekeeper is allowed six referrals is less extreme, but qualitatively similar.

Note that the referral probability for a patient of given condition varies markedly over the fiscal year. In this example, patients at or below the 46th percentile are never referred before the eleventh period. However, fortunate patients within this group may be taken in periods eleven or twelve if no high-value patients have previously arrived. In contrast, the referral probability for a patient at or above the 92nd percentile continuously declines over the fiscal year, since there is a positive probability that the gatekeeper will have previously expended the entire budget and will therefore be unable to provide this high-value referral.

For patients whose condition falls between these two extremes, the referral probability moves more irregularly over the fiscal year. For example, a 75th percentile patient will be referred in the seventh period as long as the gatekeeper has two remaining referrals. The same is true in periods eight or nine, but the gatekeeper is less likely to have two referrals to give. By period ten or later, the 75th percentile patient will be referred even if there is only one referral left. Hence as shown in the figure, the referral probability for this patient has two interior peaks.

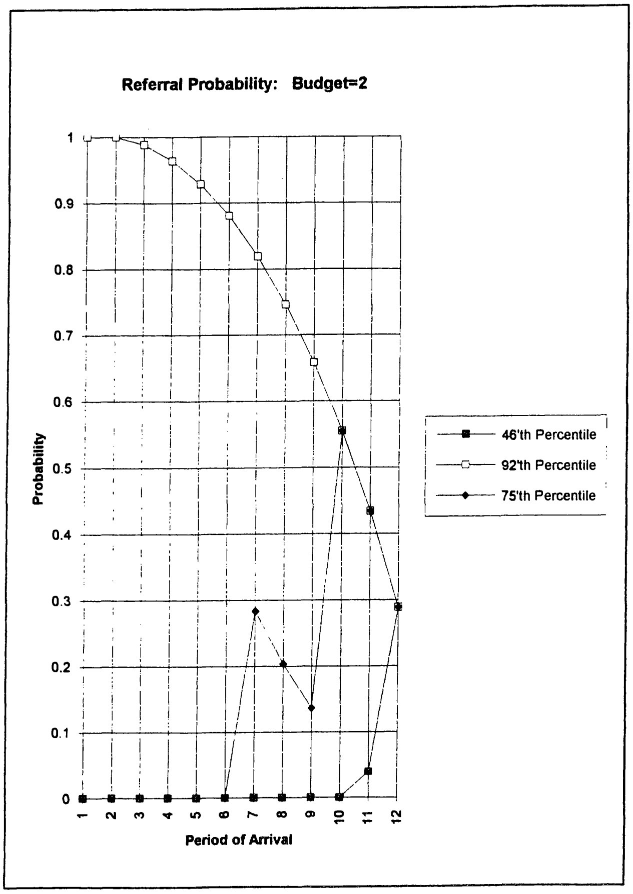
Figure 2 Mean Value of Referrals over Time



5. Two Approaches to Ameliorate the Inefficiencies of Decentralized Budgetary Expenditures

How can a central budgetary authority, who cannot examine individual patients but can observe overall costs,

Figure 3 **Referral Probability for an Individual (Various Percentiles)**



induce gatekeepers to approach socially optimal decisions? Within the framework of a decentralized organization, the best she can do is to choose budgetary amounts that will (duly accounting for the relevant incentives) maximize expected value net of budgetary costs. The inefficiencies of this process are illustrated in Figure 1B. As gatekeepers' thresholds diverge from true social cost, deadweight loss emerges. Patients who would derive benefits greater than the true social cost are sometimes turned away. Others are referred who will derive relatively little benefit when there happens to be "excess capacity" near the end of the fiscal year.

Decentralized budgets are inefficient because although each subunit learns that the social value of expenditures rises and falls over the fiscal year, such information is not monitorable from the center. Hence no desirable reallocation is made. Assuming that improved monitoring is not feasible, the most promising response is to smooth out the value of expenditures. We consider two ways this might be achieved. First, one can lengthen the fiscal year. In effect, a larger sample size moves each subunit proportionately closer to the mean. Second, one can aggregate the budgets of different subunits. Each of these strategies entails costs, of course. If we lengthen the fiscal year, we are less able to respond to whatever new information does flow to the center. If we aggregate budgets, and group two or more subunits together, the familiar common pool problem of excess expenditure is reintroduced, albeit at a lower administrative level. An appropriate policy would balance the smoothing benefits of longer or more aggregated budgets against their respective costs of lessened responsiveness and common pool excess. Many current policies, such as the overwhelming use of an annual budget without carryovers, appear to make no effort to conduct such balancing.

5.1. Lengthening the Fiscal Year

We first explore the possibility of lengthening the fiscal year. At the federal level, this is effectively accomplished for some programs using multiyear budgets and appropriations. For example, Congress sets aside multiyear procurement budgets for major defense projects.

One can also allow subunits to carry a portion of unspent monies forward into the next fiscal year. Indeed Vice President Gore (Gore 1993, p. 17) has proposed that federal agencies be allowed to carry over half their unspent budgets into the next fiscal year. Such carryovers improve efficiency by preventing the shadow price of spending from dropping to zero at the end of the fiscal year. At the same time, such carry-overs overcome only some inefficiencies; the doctor may still be overly selective early on.

A second strategy is to allow the doctor to borrow against future budgets. Notice that if one allows complete carry-overs and complete borrowing against future expenditures, in essence the doctor must maximize his expected utility over an infinite time horizon. If there is no penalty for borrowing against future budgets, one can show that the doctor's optimal strategy is to employ a constant threshold independent of time or previous expenditures. Thus, the center can induce optimal behavior by assigning a budget $b = T(1 - F(1))$—a per-patient budget equal to the probability that the benefit from referral exceeds the cost—and also allowing carryovers without penalty. Proposition 2 shows that such a policy achieves first-best outcomes with decentralization.

One way to think about the improvements that carrybacks and carry-forwards allow is to compare them to savings and credit markets in the larger economy. Just as credit markets raise individuals' lifetime utility by smoothing consumption, borrowing allows government agencies to provide a more constant social benefit stream across time.

When time is short, the physician gatekeepers diverge markedly from the social optimum. And if the time period is sufficiently long, a decentralized expenditure approaches the first-best allocation. This intuition favoring long budgetary cycles is formalized below.

PROPOSITION 2. *Suppose that the true referral cost is \$s and a fiscal year is T periods long. Moreover, suppose that the cost-effectiveness of referrals is given by the distribution F(z) and its associated density f(z).*

In the absence of discounting, if each doctor is given the break-even budget $T\Theta = T^*[1 - F(s)]$ *and if* $V(T\Theta, 0)$ *is the doctor's expected payoff following his privately optimal rule, we have*

$$\lim_{T\to\infty} \frac{V(T\theta, 0)}{T\theta} = E[Z \mid Z > s]. \tag{12}$$

PROOF. See Appendix.

Notice that from the social perspective, the expected payoff per patient is the expected benefit minus referral cost, or $[V(T\Theta, 0)/T\Theta] - s$. By Proposition 2 this converges in expected value to $E[Z \mid Z > s] - s$. This is exactly the expected payoff from the socially optimal policy. So decentralization approaches the first-best outcome as the fiscal year is lengthened.

In practice, any system of saving unspent funds or borrowing against future budgets will utilize interest rates. To achieve efficiency, expenditure patterns can be expected to readjust so the implicit discount rate across periods equals the interest rate. Proposition 2 is generally false in the presence of discounting and interest payments, even when these two rates are the same.[13] However, Monte Carlo simulations suggest that lengthening the fiscal year will achieve close to the first-best outcome.

5.2. Caveats

We identify three major problems with the long budget year: strategic excess spending, underweighting of the future, and an inability to adjust to new information.

5.2.1. Strategic Excess. Say that a college's financial aid officer had a three-year budget for expenditures, and that he spent 75 percent the first year. He would surely be fired. Yet what if he only spent 45 percent of the three-year total? He might expect the president and the trustees of the college to kick in more money for financial aid over the remaining two years, thinking that the first year was one of unusually high need. Easy cross-period transfers also create the potential for Ponzi-like behavior, with a subunit going deeper and deeper into debt "borrowing" from future budgets. The inability to distinguish between an exceptional year of high value and strategic distortion imposes a significant constraint on the budget-lengthening strategy.

In similar fashion, the prospect of a "soft" budget constraint lessens the credibility of schemes to enforce fiscal discipline.[14] This has been an especially serious constraint on the use of contingency funds. As one Congressman succinctly observed:

> [i]*t* is a matter of public knowledge that in many instances this contingency fund has been used for one contingency only and that contingency is that the House and Senate did not appropriate as much money for this program as the people downtown would like . . . (Fisher 1975, p. 67)

Kearns (1994) provides some evidence that biennial budgets lead to increased expenditures. She suggests

[13] However, given equality of discount and interest rate, if the number of arrivals also grows as the interest rate (neglecting the issue of indivisibility) then the analysis collapses to the analogous result of Proposition 2.

[14] Parents supporting ne'er-do-well children encounter similar problems. If they provide resources for the long-term (perhaps to instill financial responsibility), they may find the budget rapidly exhausted, with themselves thrust into the uncomfortable position of choosing between providing more and having a destitute child.

that agencies and aligned constituencies exploit longer budget cycles to countermand fiscal restraints through the use of earmarking and other off-budget mechanisms. Such strategic opportunities mean that lengthening the budget year requires more expensive and time-consuming monitoring. These costs must be measured against the efficiency gains that long-term budgets have the potential to provide.

5.2.2. Underweighted Future. The proof of Proposition 2 assumes that the gatekeeper's subjective discount rate is zero.[15] The policy import of these results is therefore greatest for settings where the time horizons required to smooth expenditures are relatively short compared with the relevant discount rates employed. Lengthening the fiscal year can be disastrous if impatient officials can overspend early-on during a lengthened fiscal year.

The problem is especially acute because while the lifespans of agencies or firms are very long, the working lives of most officials are not. Irresponsible officials can run up large, suboptimal debts and exit the scene before this becomes publicly known. These pressures provide the motivation for balanced-budget amendments in many states. For similar reasons, the norm among most nonprofit institutions, if only to reassure donors, is to run in the black.

The short-horizon problem is familiar within the private sector, where executives can neglect building maintenance or employee training and then depart before the resulting problems become apparent. Yet stock market analysts at least try to ferret out such concerns, and stock prices provide some measure of future problems and obligations. Information asymmetries are more troublesome in the public sector, where the press often does a poor job of monitoring, and political opponents identify problems even where they do not exist. Kotlikoff (1991) suggests that federal taxes in the year 2020 will be far higher than they are today, at least partly because information that obligations are being shifted is poorly conveyed to the public until it is too late. The proliferation within government of relatively temporary political appointees—many focused on immediate policy disputes rather than their agency's long-term reputation and efficiency—further aggravates these short-horizon problems.

5.2.3. New Information. One reason for starting budgets periodically afresh is that new information becomes available. Such information could relate to changing marginal productivities of expenditures or to the changed preferences of the central authority. Long-term budgets cannot respond to such information. Moreover, if the central authority were to change, no one would believe that the old budget would be (or should be) maintained intact. When we lengthen the budget period, we sacrifice flexibility to improve allocations.

This need for flexibility is one reason that both Congress and state legislatures have proved relatively unenthusiastic about biennial budgets. In 1940, 44 states employed biennial budgets. Only 19 states continued the practice by 1987. Uncertainties in predicting revenues and economic conditions present especially serious problems for long-term budgeting.[16]

5.3. Combining Subunits

A second strategy to combat inefficiency due to local variation is to combine different subunits. A doctor in a group practice has less need to be over-cautious early on if he can pool this risk with several partners. Such risk-pooling may bring other benefits by fostering consultation and peer monitoring. This strategy also ameliorates problems of indivisibility. If expenditures are lumpy and the optimal expected expenditure is 1.5 units per gatekeeper, there can be large efficiency gains from grouping gatekeepers together and allocating the correct expected budget.

This strategy may only work for small groups. A doctor in a small practice will not grant a needless referral if he thinks he is depleting an asset he himself will later need. Yet as group size increases, monitoring difficulties increase, and so does the individual incentive to free-ride by depleting the common pool.[17] Yet even when

[15] We thank an anonymous referee for noting the importance of this point.

[16] Kearns (1993, pp. 41–44) provides a useful summary of state practices. Meyers (1988, pp. 27–28) simulates the effect of economic uncertainty on federal revenues.

[17] The accelerated depletion problem arises with diners sharing dishes at a Chinese restaurant. They eat quickly so as to garner a larger share.

monitoring is infeasible, an individual gatekeeper still has some incentive to preserve a common resource that he himself may later need.

Consider the effects of group size on budgetary behavior, leaving aside any monitoring. Start with a simple two-person, two-period game. Two doctors are grouped together; each sees one patient per period. The doctors are allowed to make a total of two referrals between them. (In case they both want to make a referral and there is only one left, they flip a fair coin.) Ex ante, a patient has probability p of being "high-need," in which case the patient derives a benefit $x + \pi$. With probability $1 - p$, the patient is "low-need," and will only receive benefit x.

One possible Nash equilibrium has each doctor refer as many patients as he can, independent of underlying need. This is the case in which the doctors deplete their common pool, and hence gain nothing from their group arrangement. Yet as long as $p\pi > x$—that is, the option value to a doctor of one referral exceeds the benefit to a low-need patient—it is also an equilibrium for each to only refer high-need patients in the first period. Hence there is no need for monitoring, since neither doctor has an incentive to waste a referral he might want for a high-need patient in the second period.

The game can easily be extended to the case of N doctors sharing $M < N$ referrals over two periods. If $B_{N,p}$ denotes a binomial random variable with parameters N and p, the doctors will cooperate appropriately in the first period so long as

$$\frac{p\pi - (N-1)x}{x} \geq \frac{M}{p} \frac{\Pr[B_{N,p} > M]}{\Pr[B_{N-1,p} < M]}. \qquad (13)$$

To combine subunits to reduce deadweight loss from decentralization, in principle one should form the largest group for which (13) holds. In real life, however, the members of combined subunits may be able to engage in peer monitoring and sanctions to guard against overspending. If so, somewhat larger groups would be desirable.[18]

5.4. Transferring Funds

Still a third strategy, in a somewhat different spirit, has been highlighted by Vice President Gore (1993, p. 13). His "National Performance Review" would allow public agencies greater flexibility to transfer funds between different accounts. If an agency has $40,000 left over for office supplies but is short $50,000 in maintenance, it seems sensible to allow the agency to transfer the $40,000 to help cover its maintenance costs. The danger, of course, is that such transfers frustrate the wishes of the initial budgetary authority, which presumably had reasons for dividing resources between office supplies and maintenance as it did. The closer the correspondence in preferences between the center and its subunits, the broader budget classifications will be, which makes efficient reallocation easier.

In some circumstances, subunits are granted considerable general discretion, but are constrained in the administrative or policy areas where preferences are most likely to diverge. For example, most organizations (and virtually all civil service organizations) impose considerable central control on the compensation of personnel. More than equity across subunits is involved. Salary guidelines regulate the area in which subunits would most like to spend against the preferences of the center. Restrictions on hiring are also common, particularly in organizations, such as most government agencies, where layoffs are expensive. These restrictions are designed to prevent subunits from "creating facts," the need to pay a new salary into the future.[19]

The central authority should also be concerned with subunits' attempts to commit to long-term projects. Substantial down payments on indivisible or increasing-returns projects would increase the cost-effectiveness of future expenditures, hence make them more likely to be made. Another danger is represented by the ability of some expenditures, say for an innovative social service, to mobilize powerful constituencies that will oppose future cuts.[20]

[18] A referee notes that combining subunits may foster collusion against existing central authorities. The strategic consequences of combining subunits is an intriguing topic for future investigation.

[19] Fisher (1975, pp. 99–122) describes policy disputes leading to Congressional mandates to segregate line-item funds.

[20] Gaffers and Kelley (1991) describe two Minnesota cases in which this occurred. They conclude that "allowing unspent funds to be 'lost'

This illustrates the observation of McCubbins et al. (1989) that legislators are especially reluctant to relinguish ex ante control over matters that are politically difficult to correct ex post. Thus, agencies whose programs are sustained by divided political coalitions are especially likely to face stringent budgetary controls to prevent the agency and its allies from departing from the legislative equilibrium undergirding the agency's mandate.

Elected officials are also reluctant to cede control over sensitive matters that bring highly visible benefits to important constituencies. In recent years Congress has been unwilling to allow the armed services to determine base-by-base allocations for areas such as quality of life and readiness. Congress has also firmly resisted frequent recommendations that major defense procurement be undertaken on multiyear budgets.[21]

6. The Potential for Budget Innovation

Private firms often respond to incentive problems by decentralizing property rights. Contract theorists argue that if managers have access to important private information, the efficient solution is to grant them whole or partial ownership over the residual returns they effectively control. This is one reason for the trend toward breaking up conglomerates. It also provides the motivation for stock compensation plans for corporate CEOs.[22] Similar arguments have been used for capitated payment schemes that allow doctors and hospitals to increase profits by lowering their costs.

Unfortunately, such efforts to reassign property rights are not applicable to many environments in the government and nonprofit sectors. Managers cannot own stock in their government agency or in most nonprofits. And in many nonprofits, large salaries are considered unseemly, perhaps betraying a limited commitment to the organization's mission. Even within for-profit firms, there are major obstacles to reassigning property rights as a response to the problems of decentralized decision-making. The value of managerial stock options, for example, depends upon the performance of the whole firm, not just a specific manager's domain.

We have identified several measures that in principle might increase the efficiency of decentralized systems. Unfortunately, political and organizational constraints often limit their use. For example, we recommend that budgetary carryovers and borrowings should be explored to see where they can promote efficiency. While such strategies are appealing to the economist, politicians have proved conspicuously reluctant to institute innovations that might undermine their ability to control agencies' activities and expenditures. Legislatures direct executive agencies through the power of the purse. Budget allocations that last longer than a legislative term (two years for Congress and many state legislatures) seem very unlikely. The potential for carryovers is also limited because of credibility concerns. If a state agency were to end the fiscal year in surplus, there would be great pressure to return the funds to the general pool and probably to cut the agency's future funding.

Despite these difficulties, carryover systems have frequently been proposed, though at times the cynic would suggest. Such measures are usually proposed in times of deficit, when we can pretend that we are establishing a system that balances out over time.

Strategies such as the Oregon Medicaid plan represent a different approach to decentralization. In essence, these strategies seek to reduce agency losses by providing more detailed central direction. Oregon allocated Medicaid funds in accordance with a cost-effectiveness ranking for common treatments. Such a strategy avoids some shortcomings of the gatekeeping approach. It allocates more funds to doctors who serve sicker populations. It also reduces undesirable incentives for dynamic inefficiency.

Compared with gatekeeping, however, the Oregon approach suffers in two ways. First, it does not respond to local information. Not all coronary bypasses are the same, and the center cannot capture these differences. Second, it may be more vulnerable to political factors

may, indeed, be the only way to preserve the long-term financial stability and public confidence social programs need to succeed."

[21] Former Congressman Mickey Edwards, personal communication, October 7, 1993.

[22] Grossman and Hart (1986).

that encourage inefficiency. For example, Oregon was not able to use age as a treatment criterion, and the designing commissioners were permitted to make adjustments that depart from cost-effectiveness.

Any budget innovation will entail costs, including political costs. Nonetheless, strategies should be considered that could reduce the predictable inefficiencies of fixed-budget decentralization. The central planner should give considerable thought to the optimal size of subunits under her control, to exploit local risk-pooling and to capitalize on peer monitoring. Longer fiscal years, giving subunits greater incentive to practice long-term planning, could prove desirable for some classes of expenditure. Block grants, on a strict formula basis, eliminate distortionary efforts by subunits to appear more deserving.

7. Conclusion

The budgetary process involves a subtle game between the central authority and organizational subunits. Variability in the value of subunit expenditures, the focus of this paper, can be ameliorated but not eliminated, since the ameliorating measures create new difficulties of their own. This suggests a depressing meta-theorem: The asymmetries of information and expertise that make decentralization attractive inevitably create losses due to agency problems that arise once decentralization is implemented. These agency losses are of two kinds: variability in the value of expenditures when budgets are fixed over the fiscal year, and gaming to influence the budget when funding is responsive to patterns of expenditure. A desirable management strategy must strike a balance between these concerns.[23]

Appendix Proofs of Propositions 1 and 2

Proof of Proposition 1. Before proving the proposition, it is useful to state the following lemma, which reflects standard results:

Lemma 1. *Let X and Y be two random variables such that E[X] = E[Y], and X exhibits second-order stochastic dominance. That is, for all Γ,*

$$\int_{-\infty}^{\Gamma} [F_Y(z) - F_X(z)]dz \geq 0. \tag{14}$$

Then for all Γ,

$$\int_{\Gamma}^{\infty} zf_X(z)dz \leq \Gamma[F_Y(\Gamma) - F_X(\Gamma)] + \int_{\Gamma}^{\infty} zf_Y(z)dz. \tag{15}$$

Proof. Integrating (14) by parts yields that

$$\Gamma[F_Y(\Gamma) - F_X(\Gamma)] \geq \int_{-\infty}^{\Gamma} z[f_Y(z) - f_X(z)]dz. \tag{16}$$

Now X and Y have the same mean. So

$$\int_{-\infty}^{\infty} z[f_Y(z) - f_X(z)]dz = 0. \tag{17}$$

Subtracting (17) from (16) yields the lemma. □

Proof of Proposition 1. Let the distributions X and Y be as before, so that X exhibits second-order stochastic dominance. Let $V_X(b, t)$ and $V_Y(b, t)$ be the associated value functions. The proof proceeds by induction. Note that at the final period T, $V_X(b, T) = V_Y(b, T)$. The base of the induction hypothesis is therefore trivial.

Suppose further that for all $b < B$ and all t, $V_X(b, t) \leq V_Y(b, y)$. Moreover, suppose that for all $\pi > t$, $V_X(B, \pi) \leq V_Y(B, \pi)$. For the induction hypothesis to carry through, we need to show that $V_X(B, t) \leq V_Y(B, t)$.

If $C_X(b, t)$ is the set of optimal thresholds appropriate for the distribution X, an optimizing gatekeeper facing distribution Y can do *no worse* than follow the rule $C_Y(b, t) = F_Y^{-1}[F_X(C_X(b, t))]$. This means that C_Y occupies the same percentile in the Y-distribution as C_X does within the X. Writing the value functions, and exploiting the induction hypothesis $V_X(B, t + 1) \leq V_Y(B, t + 1)$, and $V_X(B - 1, t + 1) \leq V_Y(B - 1, t + 1)$, some algebra yields

$$\begin{aligned} \Delta V &\equiv V_Y(B, t) - V_X(B, t) \\ &\geq [1 - F_Y(C_Y)]E_Y[Y \mid Y > C_Y] \\ &\quad - [1 - F_X(C_X)]E_X[X \mid X > C_X] \\ &= \int_{C_Y}^{\infty} zf_Y(z)dz - \int_{C_X}^{\infty} zf_X(z)dz, \end{aligned} \tag{18}$$

$$\int_{C_Y}^{\infty} zf_Y(z)dz \geq \int_{C_Y}^{\infty} zf_X(z)dz - C_Y[F_X(C_X) - F_X(C_Y)] \tag{19}$$

by Lemma 1 and the definition of C_Y. Integrating by parts and simplifying,

$$\Delta V \geq F_X(C_X)[C_X - C_Y] - \int_{C_Y}^{C_X} F_X(z)dz. \tag{20}$$

Since $F_X(z)$ is increasing, ΔV is positive regardless of the algebraic sign of $[C_X - C_Y]$. So $V_Y(B, t) \geq V_X(B, t)$. The induction argument is thereby confirmed, and Proposition 1 follows. □

[23] John Vickers, John Lindsay, and John Pratt provided especially helpful advice. Bradford Gray, an associate editor, the referees, and participants at the October 1994 ORSA-TIMS meeting provided useful comments. The National Science Foundation and the Robert Wood Johnson Health Policy Scholars Program provided financial support.

PROOF OF PROPOSITION 2. Suppose that the true social cost of treatment is \$s. Let $\Theta = 1 - F(s)$. Moreover, define $\eta = E[Z|Z > s]$. Here η is the expected benefit to a patient who exceeds the treatment threshold \$$s$. Of course the socially optimal policy is to refer every patient who would derive more than \$$s$ from the treatment, yielding a per-patient expected social payoff from referrals of $[\eta - s]$.

We want to show that if a gatekeeper is given a budget of $T\Theta$ referrals for T periods, he will approximate the social optimum. That is, for any ϵ and for large enough T, his per-patient expected payoff is no less than $\eta - \epsilon$ when he follows the optimal rule.[24]

Before proceeding, the following facts about the binomial distribution will be useful. Let X be the random variable such that $X = 1$ with probability Θ and $X = 0$ with probability $1 - \Theta$. Then if there are T arrivals in a fiscal year, let $\{X_1 \cdots X_T\}$ be an independent, identically distributed collection of random variables with the same distribution as X. If $Y(T) = X_1 + \cdots + X_T$, Y will be binomially distributed with mean $T\Theta$ and variance $T\Theta(1 - \Theta)$. $Y(T)$ is the number of patients over the fiscal year who would derive more than \$$s$ from treatment.

For any $\epsilon > 0$ and $r = \Pr[Y < T\Theta - T\epsilon)]$, Chebyshev's inequality implies that

$$r \leq \frac{\Theta(1 - \Theta)}{\epsilon^2 T}. \tag{21}$$

Note that for any fixed ϵ, r converges to zero as $T \rightarrow \infty$.

Now suppose that a gatekeeper is given a budget of $T\Theta$ referrals for T periods. How will he spend it? Consider the following three strategies:

- S_1: Accept all referrals with benefit exceeding \$$s$ until either the budget is exhausted or the end of the fiscal year is reached.
- S_2: Accept all referrals exceeding \$$s$.
- S_3: Follow the gatekeeper's (privately) optimal rule.

S_2 is the socially optimal strategy; of course it may require exceeding the gatekeeper's available budget. From the gatekeeper's perspective, the expected payoff from S_3 at least equals the expected payoff from S_1. By the way Y was constructed, the gatekeeper's payoff must exceed $(1 - r)(T\Theta - T\epsilon)\eta$. So the per-patient expected payoff—$V(T\Theta)/T\Theta$—must exceed $(1 - r)(\Theta - \epsilon)\eta$.

Now as $T \rightarrow \infty$, for any ϵ we know that $r \rightarrow 0$. Moreover, notice that ϵ can be made arbitrarily small. So the gatekeeper's per-patient expected payoff satisfies

$$\underline{\lim} \frac{V(T\Theta)}{T\Theta} \geq \eta. \tag{22}$$

We also want to show that η provides an asymptotic upper bound as well. Consider the additional strategies S_4 and S_5 defined as follows:

- S_4: Accept the $T\Theta$ highest-value referrals.[25]
- S_5: Follow S_4 or else accept all referrals exceeding \$$s$, whichever yields the gatekeeper more.

Both strategies require omniscience and yield the gatekeeper a higher expected payoff than does his optimal feasible strategy S_3. Clearly S_5 yields the gatekeeper a higher payoff than does S_4.

But note that the per-patient expected payoff from S_5 is the same η that comes from following S_2, *except* that under strategy S_5 the gatekeeper also derives additional utility from "bonus" referrals worth \$$s$ or below.

Yet these bonus referrals yield a small expected payoff. For any $\epsilon > 0$, the probability of having ϵT or more of such bonus referrals is given by r bounded by (21) above. In that case the per-patient expected payoff from bonus referrals is at most $[rsT\Theta]/T$ which converges to 0 as $T \rightarrow \infty$. Similarly, when there are no more than ϵT bonus referrals the per-patient payoff is at most $[s\epsilon T]/T$. Since ϵ is arbitrary this also can be made arbitrarily small.

This argument implies that the gatekeeper's per-patient expected payoff satisfies

$$\overline{\lim} \frac{V(T\Theta)}{T\Theta} \leq \eta. \tag{23}$$

This is exactly what we needed to prove. □

[24] John Lindsey provided a clever simplification of the proof in a previous version.

[25] We assume $T\Theta$ is rounded down to the nearest integer. This is irrelevant to the argument for large T.

References

American Medical News, "The Good Gatekeeper," October 3, 1994, pp. 27–29.

Banks, J., "Agency Costs, Cost Information, and Auditing," *American J. Political Science*, 33 (1989), 670–99.

—— and B. Weingast, "The Political Control of Bureaucracies Under Asymmetric Information," *American J. Political Sci.*, 36 (1992), 509–524.

Bendor, J., S. Taylor, and R. Van Gaalen, "Politicians, Bureaucrats, and Asymmetric Information," *American J. Political Sci.*, 31 (1987), 796–828.

Besanko, D. and D. Sappington, *Designing Regulatory Policy with Limited Information*, Harwood Academic, New York, 1987.

Fisher, L., *Presidential Spending Power*, Princeton University Press, Princeton, NJ, 1975.

Gore, A., *From Red Tape to Results: Creating A Government That Works Better and Costs Less*, Plume Books, New York, 1993.

Grossman, S. and O. Hart, "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *J. Political Economy*, 94 (1986), 691–719.

Hayek, F. A., "The Use of Knowledge in Society," *American Economic Rev.*, 35 (1945), 519–530.

Holmstrom, B., "Moral Hazard in Teams," *Bell J. Economics*, 13 (1982), 324–340.

Ibbotson, R. and A. Sinquefield, *Stocks, Bonds, Bills, and Inflation: Historical Returns*, Research Foundation of the Institute of Chartered Financial Analysts, Charlottesville, VA, 1989.

Jefferys, M. and A. Kelley, "Program Budgeting," *Public Budgeting and Finance*, 11 (1991).

Karlin, S., and M. Taylor, *A Second Course in Stochastic Processes*, Academic Press, New York, 1981.

Kearns, P., "The Determinants of State Budget Periodicity: An Empirical Analysis," *Public Budgeting and Finance*, 13 (Spring 1993), 40–59.

——, "State Budget Periodicity: An Analysis of the Determinants and the Effects on State Spending," *J. Policy Analysis and Management*, 13 (Spring 1994), 331–362.

Kotlikoff, L., *Generational Accounting*, Free Press, New York, 1991.

Kumar, A., P. S. Ow, and M. Preitula, "Organizational Simulation and Information Systems Design," *Management Sci.*, 39, 2 (1993), 218–241.

Lipsky, M., *Street-Level Bureaucracy*, Russell Sage, 1980.

Malone, T., "Modeling Coordination in Organizations and Markets," *Management Sci.*, 33, 10 (October 1987), 1317–1332.

Marschak, J., and R. Radner, *The Economic Theory of Teams*, Yale University, New Haven, CT, 1972.

McCubbins, M., Noll, R., and Weingast, B., "Structure and Process as a Solution to the Politician's Principal-Agent Problem," *Virginia Law Rev.*, (1989), 431–482.

—— and Schwartz, "Congressional Oversight Overlooked: Police Patrols Versus Fire Alarms," *American J. Political Sci.*, 28 (1984).

Meyers, R., "Biennial Budgeting by the U.S. Congress," *Public Budgeting and Finance*, (Summer 1988), 21–32.

Milgrom, P., and J. Roberts, *Economics, Organization, and Management*, Prentice-Hall, Englewood Cliffs, NJ, 1992.

Miller, G. and T. Moe, "Bureaucrats, Legislators, and the Size of Government," *American Political Sci. Review*, 1983, 297–322.

Mondschein, S., *Optimal Fail Strategies in Stochastic, Dynamic Environments*, Unpublished Doctoral Dissertation, MIT, Cambridge, MA, 1993.

Niskanen, W., *Bureaucracy and Representative Government*, Aldine Press, Chicago, 1971.

Pratt, J. and R. Zeckhauser, *Principals and Agents: The Structure of Business*, Harvard Business Press, Cambridge, MA, 1987.

Washington Monthly, "Memo of the Month," December 1987.

Weimer, D., "Claiming Races, Boiler Contracts, Heresthetics, and Habits: Ten Concepts for Policy Design," *Policy Sciences*, 25 (1992), 135–159.

Wildavsky, A., *The Politics of the Budgetary Process*, 2nd Ed., Little Brown, Boston, 1974.

Accepted by Edward H. Kaplan; received November 16, 1993. This paper has been with the authors 8 months for 2 revisions.