Toward a More General Theory of Governmental Structure

By MANCUR OLSON*

A theory of governmental structure begins most naturally with why we need governments. Governments are not needed to perform any functions that markets perform perfectly. Thus a theory of governmental structure naturally begins with market failure. Some economists are uncomfortable with the phrase "market failure" because it is not a sufficient reason for governmental intervention; governments also fail and may perform worse than an imperfect market. This argument, though correct, does not eliminate the need for the concept of market failure. If markets never fail, there is no need for government, so a concept of a market failure is required for an adequate normative theory of government. Since most governments do not restrict themselves to correcting market failures, we cannot explain much of what governments actually do in terms of market failures. Yet we shall see that a theory of governmental structure that begins with a normative analysis of market failure also helps to explain some crucial aspects of reality. Market failure can be quite important, as is shown by the tendency of people throughout history to flee from areas of anarchy to areas with governments, bad as those governments often are. This suggests that the consumers' surplus arising from the elemental services of government is quite large.

Two well-known explanations of market failure can be generalized to cover all types of market failure. When large numbers rule out Coasian bargains, there are two conditions, each of which is sufficient but not necessary for market failure: nonrivalness in consumption and the infeasability of excluding nonpurchasers. Though these two attributes are often found together and are often considered necessary attributes of public goods, it is important for the present purposes to distinguish them.

The additional consumption of some goods, such as the further use of an idea, the additional viewing of a television program, the additional crossing of an uncongested bridge, or an additional observer of a high-wire act in an uncrowded setting, need not reduce the consumption of others. In these cases of classic nonrivalness, additional consumption has no marginal social cost, so it is a necessary condition of Pareto efficiency that the good not have a positive price. Such a good cannot be provided at the marginal social cost price of zero by a private firm that must cover its costs of production and cannot engage in price discrimination. This is true even when, as is often the case, nonpurchasers can be excluded at little or no cost. Even if some amount of the nonrival public good may sometimes be provided by a private firm that charges a monopoly price (or an average cost price), the less-than-optimal output that results from a nonzero price entails that the first-best conditions for Pareto efficiency are not met. In these cases, society faces a tradeoff between the losses from monopoly pricing and the losses from provision by imperfect governments with incentive-distorting taxation. Television is a conspicuous example of a nonrival public good subject (with scramblers) to exclusion that is sometimes privately provided at nonoptimal prices and sometimes provided by imperfect public agencies.

†*Discussants*: Rudolph Penner, U.S. Congressional Budget Office; Richard Musgrave, University of California-Santa Cruz; Gordon Tullock, George Mason University.

*Department of Economics, University of Maryland, College Park, MD 20742. I am grateful to Martin McGuire and Wallace Oates, and to the discussants of this session for most helpful criticisms, but I am solely responsible for any shortcomings.

Perhaps because nonrivalness and nonexclusion are usually assumed to go together, many economists have not noticed that nonrivalness is simply an extreme form of market failure due to economies of scale or decreasing costs. In *any* type of production in which the marginal cost is less than average cost at the marginal cost-equals-price quantity of production, there is market failure for *exactly* the same reason there is market failure for pure nonrival public goods for which exclusion is possible. Whenever the demand curve for any good cuts the marginal cost curve below the average cost curve, a firm charging the marginal cost price cannot recover its average costs without price discrimination, and thus is not viable.

At least when large numbers are at issue, markets also fail when it is uneconomic to exclude those who have not purchased a good from consumption of it. The concept of nonexclusion can also usefully be construed quite broadly. Any socially desirable redistribution of income or "social insurance" for which government is needed can be attributed to costs of exclusion. Adverse risk selection can in some cases make universal or compulsory social insurance appropriate, and this, in combination with the finding that the Friedman-Savage result on gambling is wrong and that the marginal utility of income characteristically declines with income (Martin J. Bailey et al., 1980), shows that redistributions of income to the poor that do not distort incentives too much will increase aggregate expected utility (my 1983 article). In any situation where a system of insurance has to be universal, exclusion is by definition ruled out. Just as any benefits of Richard Musgrave's (1959) "distribution branch" are a nonexclusive public good, so are any benefits of his "stabilization branch" of the government; everyone in the relevant economy tends to get the benefits of price stability and general prosperity whether he helped pay for them or not (Wallace Oates, 1972).

I. Exogenous vs. Endogenous Domains and Clienteles

The different public goods that correct different kinds of market failures have different patterns of beneficiaries. Most public goods have beneficiaries that are necessarily all within some geographical area or "domain." This domain may be either "exogenous" or "endogenous." If the domain in which the good is received is beyond the control of the political and legal system, I define it to be exogenous, but if it is determined by that system and its jurisdictional boundaries, it is endogenous to that system. Some public goods have a "clientele" that is not defined geographically. As we shall see, all public goods with clienteles rather than domains must have beneficiaries that are endogenous to the political or jurisdictional system. It will be convenient to begin with the simplest concept of the public good of exogenous domain and to assume initially that citizens do not move from residences in one jurisdiction to residences in another, but may commute anywhere within a metropolitan area.

The public good of air pollution control normally has an exogenous domain given by the relevant airshed; that of preventing a "greenhouse effect" through the burning of fossil fuels has an exogenous and worldwide domain because of the stratospheric winds. The boundaries of some other public goods and externalities of exogenous domain are given by ecological systems, by watersheds, by beautiful vistas, by the areas where accidental fires can spread from property to property, and by contagion zones (all the countries of the European continent are one zone at least for agricultural pests and diseases, but Australia is a separate zone). The boundaries of other public goods of exogenous domain are given by metropolitan boundaries, commuting distances, and social patterns that are not (except over a very long term) normally within the control of government. The boundary of the exogenous public good arising from the apprehension and incarceration of a criminal is determined by the locations of the crimes the criminal would have committed had he not been caught. Crime like many other social pathologies is mainly determined by commuting areas, so the metropolitan area is often the exogenous domain for many public goods and externalities. All public goods of exogenous domain

are *necessarily* goods for which exclusion of nonpurchasers is infeasible.

The nonexclusive public good of endogenous domain is admittedly a less-straightforward concept. It may seem that it is logically impossible to have a good for which exclusion of nonpurchasers is impossible, yet such that the government may set the boundaries within which it is received. If there were no limitations or constraints whatever limiting a jurisdiction in determining who received its services, this category would be empty. But there are constraints that give the government a range of choice in determining the geographical domain within which the public good is available, yet entail that it is nonexcludable in that area. Consider the relatively simple case of the public good of protection of property rights in real estate. A jurisdiction can protect such rights within its boundaries, but not outside them. It might seem that individuals could easily be excluded from the use of the courts and police that enforce property rights. But note that if my property rights in land in the jurisdiction are protected, and those of others in the same jurisdiction are not because the system is exclusive and not impartial, then it will not be rational for others to buy my property and it will then be worth less to me. Thus the costs of excluding individuals from the system of law and order within a jurisdiction are considerable even though the boundaries of the jurisdiction can be determined endogenously. More severe punishments by a particular jurisdiction for crimes within its legal boundaries similarly protect people nonexclusively within that jurisdiction, but may even increase crime in more lenient neighboring jurisdictions.

Though nonrivalness is often evident in nonexclusive goods, it also occurs separately. When it does, the possibility of exclusion necessarily makes the clientele endogenous. When these goods are at issue there is no need for all consumers in a given geographical area to have the same supplier, or that a given supplier restrict its supply to a given geographical area. Once produced, a television program may be viewed by any number of people in diverse areas without diminishing the value to existing consumers, but scrambling technology makes the exclusion of nonpurchasers possible. With nonrival public goods of endogenous clienteles, it is not surprising that both public and private provision occur.

New ideas that may be used by any number of people without diminishing their value to other users are nonrival public goods, even if they can be patented or copyrighted and nonpurchasers thereby excluded. New ideas of this sort characteristically arise because of both public expenditure and private entrepreneurship. When they can be effectively patented or copyrighted, they have endogenous clienteles; when they cannot, they have an exogenous worldwide domain.

There are also distinctive public goods that require separate discussion because they are relevant only in certain "catchment areas" that are determined by the pattern of preferences of the residents of the area. The patterns of preferences for public goods are dramatically dependent upon the language, religion, race, or (more generally) the culture of the people in an area. Most people obviously prefer political leaders and public servants who speak their own language, as is obvious from the histories of Belgium, Canada, Switzerland, and the successor states of the Austro-Hungarian empire. They similarly prefer governments of their own religion, ethnicity, or race, as is obvious from northern Ireland, the Middle East, and some large American cities. Over a sufficiently large number of generations, a government may create a culture that suits its convenience (France has become a country with a common language only since the end of the Middle Ages). But, in the very long run, even our grandchildren are all dead and so for practical purposes the culture of different groups must be taken as given; the taste of the Poles, for example, for distinctive and separate public goods was not erased even over several generations. When a group of people has a distinctive culture, much of the interaction, whether transfers of property rights or social pathologies, tends to be within the social group. This, along with distinctive tastes and the obvious preference for dealing with regulations in one's own language and officials of one's own culture, implies that

many of the distinctive public goods that peoples of varying cultural background want are public goods of exogenous domain.

II. Fiscal Equivalence

Whenever public goods have an exogenous domain, there are enormous advantages to government boundaries that match the exogenous boundaries of the domain of the public good. It might seem that considerations of economies and diseconomies of scale should also ideally influence the size of jurisdictions. But this is not so. The scale of the area of provision is not a variable of choice for the public good of exogenous domain. It could be that flood control has lower or higher average costs, per unit of population and property protected, with big rivers and flood plains than with small ones. But the size of the valley is given.

If the jurisdiction that provides a public good does not include all of the area in the exogenous domain of the good, there will be a spillover or external economy from the jurisdiction to neighboring jurisdictions. When the city-center government deals with air pollution or apprehends career criminals, it confers an external economy on the suburbs where some of the air pollution and criminal activity would have gone. Similarly, when Britain or the United States generate acid rain that also falls on Canada or Germany, there is again an externality that leads to less than optimal provision.

A sufficiently large government, such as a world government, would have no nonoptimality due to externalities. Many externalities are similarly avoided in large countries with unitary governments, such as France. But when the government is far larger than many of the public goods of exogenous domain, there is the political problem of the "internality" that also leads to nonoptimality. The gains from providing a local public good of exogenous domain can greatly exceed the costs of providing it, but, with a unitary national jurisdiction, the number of losers from the national taxes that would finance the public good will be far larger than the number of gainers. Thus the provision of the local public good will fail to command a majority of the larger jurisdiction. Representative political systems tend to provide some local public goods because of logrolling and other types of bargaining, but the difficulties of such bargaining are considerable and these processes often work badly. Thus there is also a grave disadvantage in having only a large unitary jurisdiction that is far larger than many of the public goods of exogenous domain.

There is accordingly much to be said for "fiscal equivalence" (see my 1969 paper), or jurisdictional boundaries that match the catchment areas of those public goods with exogenous domains. There is a need for jurisdictions that match pollution problems and other natural boundaries, and, in particular, a need for metropolis-wide jurisdictions. There is also a need for distinctive jurisdictions whenever peoples in different communities demand greatly different public goods and political leaderships. A lack of separate jurisdictions introduces a gratuitous uniformity in consumption when there are different preferences for public goods that reduces welfare (Martin McGuire, 1974). There will at the same time usually be other public goods of exogenous domain that transcend cultural boundaries, so fiscal equivalence normally calls for larger, pluralistic jurisdictions as well as smaller ones matching cultural communities. Switzerland is a country where arrangements of this kind have worked well. Canada probably would not exist at all were it not a federation with a separate province largely for Francophones.

Though fiscal equivalence calls for a multilevel mosaic of jurisdictions, it does not require separate bureaucracies or elected officials for every jurisdiction. There are economies of scope in the public as in the private sector. Often a public good of exogenous domain could be voted on (and, if chosen, paid for) only by the people in the domain, but the necessary administrative work could be handled by a jurisdiction or organization that achieved economies of scope. Some local governments apparently already provide for separate votes, tax assessments, and additional public services such as sewage or roads for distinct neighborhoods or development tracts.

In the very different case where there is endogeneity and also a clientele rather than a domain, the scale of any organization for provision should be given by the intersection of the demand curve and the marginal cost curve. With no problem of exclusion, it is an open question in each case whether there should be public or private provision. The losses from the absence of marginal cost pricing when nonrivalrous public goods are provided privately have to be weighed against the incentive-distorting effects of taxation and the imperfections of government: the market failure and the government failure need to be compared. An ideological claim that either public provision or private provision is right for all cases is unlikely to be correct. For some nonrivalrous goods, such as television, it is interesting that many countries use both public and private provision, and there is no theoretical reason why this is necessarily wrong.

With public goods from which nonpurchasers cannot be excluded but which nonetheless have an endogenous domain, the normative theory begins by suggesting that the economies and diseconomies of scale should determine the size of the jurisdiction. With governments as with private firms, there are indivisibilities in opportunity sets that give rise to economies of scale. There are also diseconomies of scale that arise because there can be only one source of coordination, and diseconomies of scale can occur when the supply of this coordination or management is combined with too large a supply of resources to be supervised effectively (Oliver Williamson, 1985). In space-intensive activities like agriculture and government, the costs of coordination are particularly great.

Consider now only local governments within metropolitan areas so that we may not only abandon the assumption that individuals do not change their jurisdiction of residence, but also suppose that such movement is costless. Let us also restrict ourselves to public goods with endogenous domains or with clienteles, so that the recipients of the public goods at issue can be determined by the political system and the sizes of jurisdictions determined by the economies and diseconomies of scale of the public good. In this set of circumstances, the famous Tiebout model applies. As Wallace Oates has helpfully pointed out to me, the movement of residence assumed in the Tiebout model effectively gives each jurisdiction an endogenous clientele, since the competitive jurisdiction can, by altering the character and level of its public good provision and the associated taxes, essentially choose its clients. As Rudolph Penner noted in his comment on this paper, the Tiebout model assumes that those with similar preferences for public goods move to the same jurisdiction, whereas this paper points out the advantages of drawing jurisdictional boundaries around pre-existing groups of similar preferences. These two approaches are complementary. The former can work well when mobility is inexpensive, as it may be within metropolitan areas over long periods, and the latter has the advantage when mobility is costly, as it is over larger areas and shorter time periods. The Tiebout model is not, however, applicable even at a local level to public goods that have a domain that is exogenous for reasons of nature or technology.

III. Military Economies of Scale

There is no reason to expect that the lowest point on the average cost curve will be the same scale for all functions of government. This consideration, like fiscal equivalence, argues against unitary governments and in favor of more differentiated governmental systems with a matrix of jurisdictions. The dramatic differences in the optimal scale of government for different functions can be illustrated by contrasting a good like protection against fires, which empirical studies have suggested can be provided efficiently on a small scale, with defense and military power. In general, an increase in the population of a country does not make it less secure. The costs of militarily defending the United States have surely not been increased because it has experienced population growth. Normally, a larger geographical area also does not add to costs of defense: Russia has been spared defeat by Napoleon and Hitler in part because it had a lot of space that permitted defense in depth and stretched

costs of defense and military power do not rise much, and may even diminish, with the population and size of a country.

Consider the per capita costs of a military capability costing, say, $200 billion. For a country of 200,000 population, the per capita cost of this military power would be an obviously unattainable $1 million per capita; for a country of 20 million, it would be $10,000, and for a country of 200 million, it would be $1,000 per head. Thus the economies of national scale in military power are staggering.

No wonder history is in large part a long story of aggression by big countries against smaller ones. The gains from exploiting the economies of governmental scale in military power are so striking that any number of kings and emperors have recognized them and used them to expand their domains.

These economies are so colossal that one must ask how there can possibly have been any equilibrium short of world government, whether arrived at by aggression, or by a peaceful Coasian agreement that would share the enormous savings among all the parties? The Napoleons of history have come close to world government; the Roman Empire, for a time, included most of the world it knew. The British Empire in the nineteenth century, though acquired in large part by classical-liberal governments interested above all in limiting government spending, came to include about one-fourth of the world's land area. So the economist must ask, why are there over a hundred independent countries in the world today and many tens of thousands of local, state, and special-purpose jurisdictions?

Part of the reason is surely the diseconomies of scale involved in coordinating and controlling vast spaces and numbers of people. But surely the advantages of something resembling fiscal equivalence have also played a role. The demand for separate jurisdictions for different cultural and linguistic groups that want distinctive and separate public goods, for separate jurisdictions that take account of other public goods of exogenous domain, and for jurisdictions that offer relief from the considerable diseconomies of scale in some local public services has been strong enough to overwhelm even the gigantic economies of national scale in military might. This suggests that there is some explanatory or predictive power in ideas of the kind that have been discussed here in a normative spirit.

This approach also suggests that the overwhelmingly large role of national governments, as opposed to both subnational and supranational jurisdictions, probably did not arise because of economies of scope or any other efficiencies. It has probably arisen mainly because national governments are the jurisdictions that have had the military or final power. This has given them the capacity to claim for themselves functions that often could have been performed more efficiently by other jurisdictions—often special-purpose jurisdictions—of both subnational and international character. Unitary national governments, in this view, are inferior to federalisms, and even federal countries could gain more from more decentralization and also stronger institutions for international cooperation.

REFERENCES

Bailey, Martin J., Olson, Mancur and Wonnacott, Paul, "The Marginal Utility of Income Does Not Increase: Borrowing, Lending, and Friedman-Savage Gambles," *American Economic Review*, June 1980, *70*, 372–79.

McGuire, Martin, "Group Segregation and Optimal Jurisdictions," *Journal of Political Economy*, January-February 1974, *82*, 112–32.

Musgrave, Richard, *The Theory of Public Finance*, New York: McGraw Hill, 1959.

Oates, Wallace, *Fiscal Federalism*, New York: Harcourt Brace, 1972.

Olson, Mancur, "The Principle of Fiscal Equivalence," *American Economic Review Proceedings*, May 1969, *59*, 479–87.

_____, "A Less Ideological Way of Deciding How Much Should Be Given to the Poor," *Daedalus*, Fall 1983, 217–36.

_____, "Space, Agriculture, and Organization," *American Journal of Agricultural Economics*, forthcoming.

Williamson, Oliver, *The Economic Institutions of Capitalism*, New York: Free Press, 1985.