The Economic Journal, **116** (*January*), 101–127. © Royal Economic Society 2006. Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

DECENTRALISATION AND ACCOUNTABILITY IN INFRASTRUCTURE DELIVERY IN DEVELOPING COUNTRIES*

Pranab Bardhan and Dilip Mookherjee

Many developing countries are experimenting with decentralisation of public service delivery to elected local governments instead of bureaucrats appointed by a central government. We study the resulting implications in a theoretical model in which the central government is uninformed about local need and unable to monitor service allocations. Bureaucrats charge bribes for services as monopoly providers, resulting in underprovision of services, especially for the poor. Local governments are directly responsive to their citizens needs but may be subject to capture by elites. Effects of decentralisation on service volumes, efficiency and equity are analysed under different financing arrangements for local governments.

The theme of the 2004 World Development Report is summarised by its opening paragraph:

> Too often, services fail poor people – in access, in quantity, in quality. But the fact that there are strong examples where services do work means governments and citizens can do better. How? By putting poor people at the center of service provision: by enabling them to monitor and discipline service providers, by amplifying their voice in policy-making, and by strengthening the incentives for providers to serve the poor.

Problems of accountability associated with traditional modes of delivery involving centralised bureaucracies include cost padding, service diversion, limited responsiveness to local needs, limited access and high prices charged especially to the poor.[1] Many developing countries have thus begun to experiment with initiatives to increase accountability of service providers by providing greater control rights to citizen groups. These include decentralisation of service delivery to local governments, community participation, direct transfers to households and contracting out delivery to private providers and NGOs. The programmes include a wide range of infrastructure services (water, sanitation, electricity, telecommunications, roads) and social services (education, health and welfare programmes). Countries where such trends have gathered

* This article extends the analysis in Part 2 of our earlier working paper titled ‘Expenditure Decentralisation and Delivery of Public Services in Developing Countries’. For financial support we are grateful to the MacArthur Foundation. Mookherjee’s research was additionally funded by National Science Foundation Grant No SBS-9709254. For helpful comments we are grateful to participants at the 1998 meeting of the MacArthur Research Network on Inequality and Economic Performance, and conferences at UC Irvine in February 2001 and at Yale in March 2001. We also thank Christopher Bliss and two referees for useful suggestions.

[1] Analysis, examples and empirical evidence concerning such ‘leakages’ and ‘targeting failures’ are provided by Banerjee (1997), Bardhan (1996), Besley (1989), Besley and Kanbur (1993), Bird (1995), Dreze and Saran (1995), Grosh (1991, 1995), Lipton and Ravallion (1995), van de Walle and Nead (1995) and the 1990, 1994, 1997 and 2004 World Development Reports.

momentum in the past two decades span different continents: Latin America (e.g., Bolivia, Brazil, Colombia, Costa Rica), Africa (Ghana, Uganda, South Africa) and Asia (Bangladesh, Indonesia, India, Pakistan).[2]

These trends towards decentralisation are difficult to interpret within the confines of the traditional literature on fiscal federalism, owing to the lack of attention devoted in that literature to problems of accountability in service delivery.[3] Instead the main focus has been the trade-off between uniformity of service provision under centralisation with problems of uneconomic scale and cross-regional externalities under decentralisation. As noted by many recent authors, the assumption of uniform provision under centralisation is neither empirically realistic, nor well founded theoretically.[4] Problems of externalities with regard to local taxation stressed by the traditional literature are also not a practical concern, given that the nature of decentralisation in most developing countries has generally taken the form of delegation of service delivery systems to local governments, without an accompanying devolution of financing authority.[5] Many of the programmes involve relatively few interjurisdictional spillovers, e.g., local water and sanitation projects. Economies of scale if significant tend to matter only with respect to production rather than distribution. Accordingly production can continue to be centralised in a public or private utility company, from whom local governments procure the service and decide how to allocate it within their respective communities. Under these conditions, decentralisation of service delivery to local governments does not involve any of the welfare losses stressed by the traditional literature.

Instead, the major concerns frequently expressed with decentralisation are that local democracy may not function appropriately, thus limiting accountability of local government officials or community leaders; see Bardhan (1996), Crook and Manor (1998), Lieten (1996), Mathew and Nayak (1996), Prud'homme (1995), Tanzi (1996), Manor (1999). With limited political contestability of local elections, leaders may be susceptible to capture by special interest groups, slacken effort to improve public services, or be incompetent, without facing any risk of losing their positions. In that case accountability, efficiency and equity in service delivery may worsen under decentralisation.

A new analytical framework is therefore needed to appraise conflicting claims about relative accountability under centralised and decentralised delivery mechanisms, which helps identify key parameters that determine the growth, equity, and welfare impact of decentralisation. In this article we provide a theoretical analysis of delivery of an infrastructural service such as water or electricity, which is entrusted either to a centralised bureaucracy or to elected

[2] For further details, see the World Development Reports of 1994 and 2004, Estache (1995), Litvack *et al.* (1998) and Bardhan and Mookherjee (2005).

[3] This literature is summarised in Cremer *et al.* (1995), Oates (1972), Musgrave and Musgrave (1984 ch. 24) and Inman and Rubinfeld (1996, 1997).

[4] See Bardhan (2002), Besley and Coate (2003), Bolton and Roland (1997), Laffont and Pouyet (2000), Lockwood (1998), Seabright (1996) and Tommasi and Weinschelbaum (1999).

[5] See, for instance, Dillinger (1995), or the various case studies in Bardhan and Mookherjee (2005).

local government officials.[6] We abstract from problems of interjurisdictional spillovers, and model accountability problems in either centralised or decentralised regimes as arising from agency problems intrinsic to either regime.

In either system we assume that service delivery has to be delegated by the central government, either to a bureaucracy or to local governments.[7] In the case of the bureaucratic system the source of the accountability problem is that the actions of bureaucrats cannot be monitored by the central government that appoints them. This is owing to high costs of communication between local areas and the central government, and difficulty faced by the central government in carrying out audits of actual service delivery patterns in local areas.[8] The bureaucrats are thus able to extract bribes from customers in their role as monopoly providers of an essential service. The centralised system ends up differentiating services to different categories of customers based on their willingness to pay bribes, resulting in non-uniform delivery patterns. However, the bureaucrats are unable to engage in perfect bribe discrimination, so the centralised system gives rise to monopoly distortions, resulting in loss of efficiency and equity. These distortions are further magnified if bribes percolate upward through multiple hierarchical layers in the bureaucracy.[9]

Decentralisation shifts control rights over service distribution to a local government subject to electoral pressure from residents. There are two classes of local users: large and small, where large users value the service more. Large users constitute a local elite commanding a higher political welfare weight, representing their superior capacity to form a special interest group, in the manner represented in capture models of Baron (1994), Grossman and Helpman (1996), Bardhan and Mookherjee (1999, 2000*a*). The extent of local capture is taken to be an exogenous parameter which summarises the effect of socio-economic inequality and political tradition within each community.

One of the main lessons of our analysis is that the effect of a switch from centralisation to decentralisation depends on the mechanism by which service provision by local governments are financed. We examine the trade-off under three different financing arrangements:

[6] The case of welfare services is somewhat different and is addressed in a separate paper (Bardhan and Mookherjee 2000*b*). In that context the service can be resold among customers, and there are divergences between ability of the poor to pay for the service and the social valuation of services delivered to them. In this article we abstract from these issues, resulting in a very different set of policy options and results. For instance it is not feasible to charge user fees for an anti-poverty programme, whereas it is feasible in the context of an infrastructural or farm input service.

[7] From the standpoint of traditional fiscal federalism this may be viewed not as a choice between centralisation and decentralisation, but between two modes of federalism: bureaucratic and political. This is partly a semantic issue; we have no major quibble with such a characterisation. However we believe that the policy trends in many developing countries described at the beginning of this article do correspond to such a choice. And the two alternatives represent varying degrees of decentralisation of monitoring and evaluation of the agents delegated responsibility over service delivery.

[8] See Wade (1997) for case studies of irrigation delivery systems in India and Korea, where the role of such problems of monitoring performance on accountability of service delivery agents is highlighted.

[9] Wade (1985) provides a vivid description of this phenomenon in the context of the irrigation bureaucracy in the Indian state of Andhra Pradesh.

(*i*) complete fiscal autonomy for local governments involving unrestricted local taxation;
(*ii*) local financing authority restricted to user fees; and
(*iii*) absence of any local revenue raising ability, wherein local governments are entirely financed by fiscal grants from the central government.

Our principal results are as follows. Under the first financing arrangement with complete fiscal autonomy for local governments, decentralisation results in an expansion of service levels provided to all categories of users. This 'pro-growth' outcome is however accompanied by increased inequality within the community, with small users bearing the fiscal burden of provision to large users. The extent of such regressive transfers depends on the extent of local capture. Moreover, rich users tend to be over-provided the service, as a result of their free-riding on the taxes paid by poor users. From a welfare point of view the effect of decentralisation is therefore ambiguous and depends on the extent of local capture.

When local governments are restricted to financing services from user fees, the scope for regressive transfers and service overprovision to the rich is correspondingly limited. In terms of both efficiency and equity, the outcome thus dominates the case of tax financing. Compared with centralisation, the outcome still involves higher service volumes for both categories of users. Moreover, it turns out that *the outcome is Pareto superior (from the citizen point of view), irrespective of the extent of local capture.* However all the welfare gains accrue to the rich, as the poor are left just as well off as under centralisation.

Finally, in the case of local services funded by fiscal grants, incentive compatibility constraints in centre-local relations cause grants to be restricted and unresponsive to local need shocks. Local governments operate under financial constraints, causing service levels to be lower compared with self-financing via taxes or user fees. At the same time the financial constraint on local governments implies that services provided to local elites come at the expense of alternative services valued by these elites, rather than levies imposed on the poor. This limits the incentive of local governments to over-provide the service to the elites. Grant financing thus results in a more equitable targeting pattern, relative to local tax finance or user fee mechanisms. In terms of overall welfare impact this has to be traded off against the lower flexibility of service levels to local need shocks. The resulting welfare comparison with other modes of financing or with centralisation ultimately depends on various parameters.

In summary, our model provides both empirically testable predictions concerning comparisons of service levels and the burden of their financing across centralised and decentralised delivery modes, as well as normative interpretations of these shifts. A general prediction is that decentralisation tends to expand service delivery levels when local governments are self-financing and the effect is greater when they have greater fiscal autonomy. This is exactly what Estache and Sinha (1995) find in a study of 20 countries over the period 1970–92. At the same time our model suggests that it would be a mistake to infer from this that greater devolution of financing authority to local governments is desirable from a welfare standpoint, since this depends on the extent of capture of local governments.

When capture is severe, decentralisation expands service volumes the most, but imposes the fiscal burden of this on the poor. In particular, service volumes do not represent a reliable indicator of welfare impact, because they ignore the associated financing burdens.

Our model also predicts that decentralisation is associated with lower corruption as measured by bribes charged by government officials. Fisman and Gatti (2002) find this pattern in a large cross-section of countries covering the period 1980–95. Yet our model cautions against using bribes as a measure of welfare. The bribes associated with the centralised bureaucracy do disappear under decentralisation but are replaced by political influence of local elites, represented by regressive cross-subsidies hidden in government finances. Corruption measures based on bribes alone ignore political forms of corruption that may be equally or more important.

In addition, our model provides a useful way to appraise the welfare ranking of different service delivery mechanisms. User-fee financed decentralisation generally dominates both centralisation, and decentralisation with unrestricted local fiscal autonomy. For policy makers, the relevant choice should be between decentralisation financed by user fees and fiscal grants. It should be noted however, that these welfare results pertain to delivery of services to users who have the ability to pay for those services. They do not apply to poverty alleviation programmes, for instance, a context we have analysed in a different paper (Bardhan and Mookherjee, 2000*b*).

The article is organised as follows. Section 1 introduces the model, Section 2 describes the centralised regime and Section 3 the various decentralisation regimes. Section 4 discusses related literature and Section 5 concludes.

1. The Model

The service is a private benefit such as irrigation or electricity, whose production is subject to a large fixed cost F, in addition to variable costs. Production is concentrated in a single large utility in both centralisation and decentralisation regimes. There are a number of different communities, denoted $i = 1, \dots, n$. The variable cost of generating supply Y_i to community i is $\theta_i Y_i$, so θ_i is the (constant) marginal cost of delivery to community i. The realisation of θ_i is random, represented by a positive density function g_i over the interval $[\underline{\theta}_i, \bar{\theta}_i]$.

Community i has N_i users, who belong to either of two groups, large (l) and small (s), who differ in their valuation of the service. A fraction β_i of citizens in the community are small users; the rest are large users. Large users belong to a wealthy elite and value the service more owing to complementarities with other assets (land or factories) they own. The utility function of a member of group $k = l,s$ in community i is $\gamma_k \eta_i v(y_k) - t_k$, where γ_k is a group-specific valuation parameter satisfying $\gamma_l > \gamma_s > 0$, η_i is a community-specific need shock, y_k is the level of service delivered, and t_k is the net financial burden imposed on the user. The utility function v is homothetic: $v(y) = y^{\alpha+1}/(\alpha + 1)$ where $\alpha < 0$ and different from -1. Local need η_i is distributed independently across regions; within region i it has a positive density function h_i on an interval $[0, \eta_u]$ which satisfies a standard monotone hazard rate condition that $(1 - H_i)/h_i$ is non-increasing, where H_i denotes the corresponding distribution function.[10]

The central government knows only the demographic profile of the different communities, i.e., the populations N_i and its composition among the two groups β_i. It does not observe the realisation of local need or cost shocks; nor can it monitor local service delivery patterns. Owing to this lack of information, it delegates control over allocation of this service between and within communities either to bureaucrats they appoint directly (the centralised regime), or to local governments (the decentralised regime).

The fixed cost F of the utility producing the service is financed by the central government out of central taxes in both regimes; accordingly we can ignore the costs of such finance when comparing the two regimes, and focus on how variable costs are financed.[11] In the centralised system variable costs are financed by a combination of user fees and subsidies financed out of central taxes. The user fees are exogenously set at a level that is insufficient to cover operating costs, necessitating a subsidy from the central government to the utility.[12] Since central government officials cannot monitor service deliveries, the budgetary support C provided by the government to the utility must be a lump sum amount, large enough to cover its operating costs in all circumstances. Since the user fees play no role in the analysis we can set them equal to zero, without any loss of generality.[13] In the decentralised system by contrast, operating costs are funded either by local goverments out of local taxes, user fees or fiscal grants.

Central taxes involve a deadweight cost of $\lambda > 0$. Taking these deadweight costs into account, the second-best service allocation y_k^b to a group k user solves

$$\gamma_k \eta_i v'(y_k^b) - (1 + \lambda)\theta_i = 0. \tag{1}$$

The corresponding *first-best allocation* corresponds to the case where the deadweight costs of finance λ equals zero. The distributional burden of central taxes plays no role in the analysis except in the case where decentralisation is funded by fiscal grants.

2. Centralised Bureaucracy

Under centralisation, authority over service delivery is delegated to bureaucrats appointed by the central government. The bureaucracy consists of two layers. The top layer is in charge of the central utility and allocates services across communities, i.e., decides Y_i. The bottom layer consists of a bureaucrat in each community

[10] The assumption of independence of need or cost shocks across communities is a purely simplifying assumption. It only complicates the expressions for the way that intercommunity allocations react to these need and cost shocks, without altering any qualitative results.

[11] The analysis extends straightforwardly to the case where decentralisation is accompanied by privatisation of the utility, which is regulated effectively so that local governments procure the service at its true marginal cost from the utility.

[12] See Ahluwalia (1998) for a description of chronic financial problems of state electricity boards in India, resulting primarily from low levels of user fees.

[13] If l_i is a constant per-unit user fee imposed on services delivered to community i, then total user fees collected $\Sigma_i l_i Y_i$ will supplement central government revenues, thus reducing the net operating subsidy to the utility to $C - \Sigma_i l_i Y_i$.

i, who allocates Y_i across users within the community. Top layer bureaucrats observe the realisation of cost θ_i for each community i, but not the local need η_i. The lower level bureaucrat assigned to the community observes the realisation of η_i. Despite the fact that users are legally entitled to receive the service free (or against payment of the mandated user fees), the local bureaucrat will be able to charge supplementary bribes as a precondition for service delivery. We first explain how bribes are set at the local level, and subsequently how they percolate upward to higher tiers of the bureaucracy.

Consider a local bureaucrat who controls the allocation of a given aggregate service Y_i in region i among its residents. Suppose that the realisation of local need and delivery cost shocks is (η_i, θ_i). The bureaucrat cannot engage in perfect price discrimination owing to his inability to identify the precise type of any given customer. Attempts to charge higher bribes to the large users that value the service more can be circumvented by these users by masquerading as a collection of small users (e.g., by splitting their lands and assets among different family members). Given absence of resale across users, the bureaucrat's problem is to select an optimal schedule of nonlinear bribes $b(y)$, which both categories of users are subject to. Given this schedule, each user type k will decide how much service y_k to procure by maximising $\gamma_k\eta_i v(y_k) - b(y_k)$.

Using standard methods of solving such nonlinear pricing problems (Laffont and Tirole, 1993), this can be simplified as follows. Bribe and service levels for the two classes (denoted by b_k and y_k respectively) will be set to maximise (per capita) bribe income $\beta_i b_s + (1 - \beta_i) b_l$, subject to voluntary participation constraint for each class k: $\gamma_k\eta_i v(y_k) - b_k \geq 0$, the incentive constraint that large users do not seek to masquerade as small users:[14] $\gamma_l\eta_i v(y_l) - b_l \geq \gamma_l\eta_i v(y_s) - b_s$, and the allocation constraint $\beta_i y_s + (1 - \beta_i) y_l \leq Y_i/N_i$. Standard arguments can be employed to show that the participation constraint binds for small users, and so does the incentive constraint for large users, implying that

$$b_s = \gamma_s\eta_i v(y_s), b_l = \gamma_l\eta_i v(y_l) - (\gamma_l - \gamma_s)\eta_i v(y_s). \tag{2}$$

This generates the following reduced form expression for bribe income as a function of service delivery levels:

$$\beta_i D_s\eta_i v(y_s) + (1 - \beta_i) D_l\eta_i v(y_l) \tag{3}$$

where $D_s \equiv \gamma_s - [(1 - \beta_i)(\gamma_l - \gamma_s)/\beta_i]$ and $D_l \equiv \gamma_l$ represent the 'virtual' valuation parameters for the two classes respectively. Maximising bribe income (3) less variable delivery cost yields expression (7) for the intra-community service allocation given below in the statement of Proposition 1. Moreover, the local bureaucrat ends up with a total bribe income from this community of

$$N_i B_i\eta_i Y_i^{-1/\alpha}/(\alpha + 1) \tag{4}$$

where $B_i \equiv [\beta_i D_s^{-\frac{1}{\alpha}} + (1 - \beta_i) D_l^{-\frac{1}{\alpha}}]^{-\alpha} < 1$.

[14] It is well known that the solution will automatically satisfy the reverse incentive constraint as well, so small users will not try to masquerade as large users either. Moreover, the solution will automatically have the property that no large user will seek to masquerade as m (>1) small users, since the large users will be charged a discounted bribe rate relative to small users.

In the case where the central government sets user fees for financing the service, the optimal allocation is the same as long as the user fee is less than the optimal bribe. In that case the local bureaucrat charges a bribe over and above the user fee, and the absence of income effects implies that the same solution obtains (with b_s, b_l now representing the total amount paid by each type of user, the sum of the user fee and the bribe). The bribe income of the bureaucrat is correspondingly reduced by the user fees they have to remit to upper levels of the government. At the margin the bribe income of the bureaucrat is unaffected and, therefore, also their incentives.

Turn now to the allocation of service levels across communities by higher level bureaucrats. These bureaucrats will seek to extract bribe kickbacks from lower level bureaucrats in exchange for allocating service levels to their respective communities. However their lack of knowledge of local need η_i implies that they do not know how much bribe income (4) can be earned by a lower level bureaucrat from a given service level Y_i. Hence they design a nonlinear kickback schedule $Q_i(Y_i)$ specifying the kickback they demand from a lower level bureaucrat in exchange for a given service allocation Y_i.

The optimal kickback schedule can be solved as follows. Applying the Revelation Principle, the problem can be posed as follows. The local bureaucrat makes a report η_i of the local need parameter defining the bribe potential in community *i*. Following such a report, the required kickback and allocated service level is $Q_i(\eta_i),\ Y_i(\eta_i)$. The central bureaucrats select these mechanisms, one for each local bureaucrat, to maximise their own surplus, which equals the expected value of sum of budgetary slack $C - F - \Sigma_i\theta_i Y_i$ and bribe kickbacks $\Sigma_i Q_i$. Hence they design the mechanism to maximise the difference between aggregate kickback and operating costs

$$\mathcal{E}_{\eta_i}\sum_i[Q_i(\eta_i) - \theta_i Y_i(\eta_i)]. \tag{5}$$

where $\mathcal{E}_{\eta_i}$ denotes the expectation operator with respect to η_i.

At the same time the local bureaucrat is motivated to maximise the difference between local bribe income and the kickback that needs to be paid to their bosses. Hence the maximisation (5) is subject to breakeven and truthful reporting constraints for each local bureaucrat:

$$N_i B_i \eta_i v[Y_i(\eta_i)] - Q_i(\eta_i) \geq 0$$

$$\eta_i \in \underset{\tilde{\eta}_i}{\operatorname{argmax}}\{N_i B_i \eta_i v[Y_i(\tilde{\eta}_i)] - Q_i(\tilde{\eta}_i)\}.$$

Again standard techniques of solving these principal–agent problems, e.g., based on Baron and Myerson (1982), can be employed to show that the equilibrium intercommunity service allocation $Y_i(\eta_i, \theta_i)$ maximises

$$\sum_i[N_i B_i J_i(\eta_i)v(Y_i) - \theta_i Y_i], \tag{6}$$

where $J_i(\eta_i)$ denotes $\eta_i - [1 - H_i(\eta_i)]/h_i(\eta_i)$.

We thus obtain the following outcome under centralisation.

Proposition 1. *The centralised system results in the following allocation for any given state* (θ_i, η_i), $i = 1, 2, \ldots$

(*i*) *In any given state* $\{(\theta_i, \eta_i)\}_{i=1,\ldots,m}$, *the community service allocation* $Y_i(\eta_i, \theta_i)$ *maximises* (6), *resulting in underprovision relative to the first-best.*

(*ii*) *Intracommunity allocation (given the per capita service level* Y_i*) for community i is given by*

$$\begin{aligned} y_s^* &= Y_i D_s^{-\frac{1}{\alpha}}[\beta_i D_s^{-\frac{1}{\alpha}} + (1-\beta_i) D_l^{-\frac{1}{\alpha}}]^{-1} \\ y_l^* &= Y_i D_l^{-\frac{1}{\alpha}}[\beta_i D_s^{-\frac{1}{\alpha}} + (1-\beta_i) D_l^{-\frac{1}{\alpha}}]^{-1} \end{aligned} \tag{7}$$

and results in further underprovision to small users. Small users obtain a net utility of 0, while large users obtain a positive surplus, with the bribes given by (2).

Competition for rents across different layers of the bureaucracy causes the intercommunity allocation to be skewed in favour of communities with high need. Lower level bureaucrats are tempted to understate the bribe potential for their community in order to limit the kickback they have to pay their superiors. This temptation is counteracted by underproviding the service to a community when a low η_i is reported. This distortion compounds the distortion resulting from inability of local bureaucrats to price discriminate perfectly. The end result is

(*i*) underprovision of service levels to each community (relative to the first-best allocation that corresponds to a zero deadweight cost of taxes),[15] and

(*ii*) the intracommunity distortion whereby service delivery is underprovided to non-elites.[16]

It should also be noted that the statement about surplus obtained by different categories of users in Proposition 1 does not incorporate the cost of the taxes they have paid to the central government; if these are additionally incorporated then their utilities are even lower by extent that depends on the distributional incidence of central taxes.

3. Decentralisation

Now suppose authority over service delivery is devolved to local governments. They procure the service from the central utility and allocate it across local users. In order to focus on considerations related to local capture (rather than the possibility of limited technical or administrative competence of local government

[15] With a positive deadweight cost, whether there is under or over-provision depends on how large λ is.

[16] If central user fees are imposed at a constant rate of l_i for communty i, then it is easy to check that the term for costs in (6) will be modified to $(l_i + \theta_i) Y_i$, which will further compound the underprovision of the service under centralisation.

officials that might raise costs under decentralisation), we assume that they know the delivery cost θ_i and procure Y_i at this cost from the central utility.[17]

The local government may be captured by local elites owing to a variety of distortions in the functioning of local democracy. We represent the objective of the local government in community i by

$$W_i^l = \beta_i U_{si} + \delta_i^l (1 - \beta_i) U_{li} \tag{8}$$

where $\delta_i^l > 1$ represents the premium placed on the welfare of elites relative to non-elites. The switch from centralisation to decentralisation shifts control rights away from bribe extractors to those who respond to the interests of local users, owing to electoral pressures. However, they respond with a bias in favour of local elites. This bias reflects inequality within the community with regard to their wealth, literacy, social status, connections, political awareness, control over media or force. Nevertheless some degree of responsiveness of local governments to the interests of small users arises from the fact that these users often form a sizeable vote block in local elections. A local government that rides roughshod over their interests may be ejected from office by disgruntled voters. Accordingly, the degree of capture may depend on β_i, the demographic weight of small users within the community. We impose no particular structure on the capture coefficient δ_i^l, as it summarises a multitude of political determinants of local capture that we take to be exogenous.

One particular model of electoral competition that generates an objective function exactly of the form (8) is the Baron (1994) or Grossman and Helpman (1996) theory of special interest groups that contribute to campaign finance of two parties or candidates engaging in Downsian competition for local office. This version is elaborated further in Bardhan and Mookherjee (1999, 2000*a*). In that version it turns out that the extent of capture δ_i^l is an increasing function of β_i, owing to the lower level of political awareness among small users, which increases the value of campaign funds in winning elections (thus increasing the influence of elites arising from their campaign contributions). Moreover, δ_i^l tends to 0 as β_i tends to 0 and to a finite limit as β_i tends to 1. While these assumptions are inessential to our results, the Figures illustrating the service deliveries under different regimes will be drawn corresponding to such a case.

3.1. *Local-tax-financed Decentralisation*

In this version, expenditure decentralisation is accompanied by devolution of local revenue raising authority to local governments, which are fiscally autonomous and self-sufficient. Local governments have the ability and constitutional authority to finance their expenditure needs from local taxes, at the same deadweight cost λ as the central government.

Decentralisation has the advantage of exploiting the information and control possessed at the local level concerning service deliveries. On the other hand it is

[17] In the absence of this assumption, decentralisation will be subject to an additional disadvantage relative to centralisation.

subject to political favouritism by elected officials towards local elites. This takes the form of preferential service deliveries and undertaxation of the large users. The undertaxation may be achieved by selectively allowing large users to evade their tax obligations, or by designing a regressive system of local taxes (e.g., based on indirect taxes rather than property taxes). A local government will set service levels and local taxes y_k, t_k for the two classes $k = s$, l to maximise

$$\beta_i[\gamma_s\eta_i v(y_s) - t_s] + \delta_i^l(1 - \beta_i)[\gamma_l\eta_i v(y_l) - t_l] \tag{9}$$

subject to the budget constraint $\beta_i t_s + (1 - \beta_i)t_l = (1 + \lambda)\theta_i[\beta_i y_s + (1 - \beta_i)y_l]$, and non-negativity constraints on t_l, t_s.[18] Voluntary participation constraints do not need to be imposed, as local governments have the ability to impose coercive taxes and citizens are assumed unable to move across districts owing to high mobility costs. The resulting outcome is described below, and the resulting welfare compared with centralisation.[19]

Proposition 2. *Consider any state η_i such that*

$$1 + \lambda = \eta_i/J(\eta_i). \tag{10}$$

Then local-tax-financed decentralisation generates the following outcomes:

(*i*) *tax burdens $t_l = 0$, $t_s = (1 + \lambda)\theta_i[y_s + y_l(1 - \beta_i)/\beta_i]$, and service deliveries satisfying*

$$\eta_i\gamma_s v'(y_s^d) = (1 + \lambda)\theta_i; \eta_i\gamma_l v'(y_l^d) = [(1 + \lambda)\theta_i]/\delta_i^l \tag{11}$$

i.e., second-best supply to small users and overprovision to large users, implying that service levels are larger than under centralisation for both groups;

(*ii*) *higher welfare than centralisation as δ_i^l approaches* 1;

(*iii*) *lower welfare than centralisation if δ_i^l is sufficiently large.*

Assumption (10) enables us to control for the aggregate service level to the community and focus on differences in intracommunity allocations between centralisation and decentralisation. Inspecting the objective function (9) of the local government, it is evident that small users will bear the entire financial burden of the service, as local elites use their political clout to evade all tax obligations. Incorporating this financing pattern, the objective of the local governments can be expressed as a function only of the service deliveries:

$$\eta_i\gamma_s v'(y_s^d) = (1 + \lambda)\theta_i; \eta_i\gamma_l v'(y_l^d) = [(1 + \lambda)\theta_i]/\delta_i^l \tag{12}$$

[18] The non-negativity constraints prevent elites from using the local fiscal mechanism to capture the wealth of non-elites directly. Such forms of redistribution would typically be illegal. Hence reverse redistribution must be carried out indirectly in the form of distorted patterns of service delivery and selective tax evasion by elites, rather than direct transfers. If some degree of direct transfers is admitted, the outcome would be less equitable than represented below, but would not affect service levels. Consequently this version of decentralisation would perform worse relative to the other two financing variants, further reinforcing our results concerning their relative ranking.

[19] We use a utilitarian welfare criterion. However, the same results would apply with any other individualistic inequality-averse social welfare function, since the ranking of different regimes on efficiency and equity dimensions turn out to coincide.

from which result (*i*) follows. Decentralisation is thus characterised by a regressive pattern of cross-subsidisation: non-elites pay for their service as well as of the elites. Moreover, elites tend to be overprovided the service (relative to second-best cost). The greater the capture of local government the more extreme these misallocations, with lower efficiency and equity. On the other hand, with sufficiently low capture the local government maximises welfare, so the allocation approaches the second-best. Hence the welfare comparison with centralisation depends upon the extent of local capture.[20]

Nevertheless, irrespective of the degree of capture, note that that our model predicts that decentralisation expands the volume of infrastructural service delivered (assuming (10) holds). This is consistent with the empirical finding of Estache and Sinha (1995) in a cross-country context that expenditure decentralisation results in increased supply of infrastructure services when accompanied by revenue decentralisation. The principal reason for this in our model is the removal of monopoly (bribe) distortions inherent in the centralised system. Figure 1 depicts the service allocation patterns under the two systems across regions of varying demographic composition (corresponding to an increasing capture function $\delta_i^l(\beta_i)$ of the form predicted by the Grossman-Helpman model).

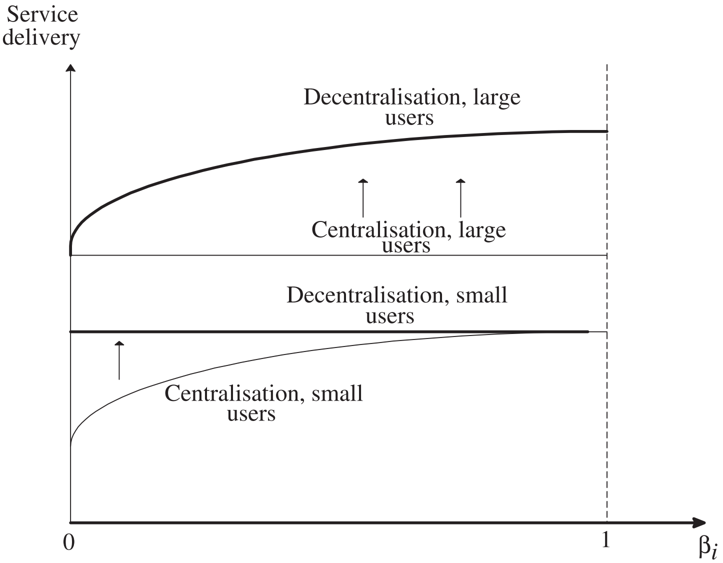

Fig. 1. *Service Delivery Patterns Under Centralisation and Local-tax-financed Decentralisation:* $1 + \lambda = \frac{\eta_i}{J(\eta_i)}$

[20] The results of Proposition 2 survive even when we include the costs of central taxes paid by users in the centralised system. Parts (*ii*) and (*iii*) are limiting results that continue to apply. They are also unaffected if provision in the centralised system is partially funded by user fees, since the service allocations and welfare calculations for different users in the centralised system remain unchanged in that case.

3.2. *User Fee Financing*

In practice, local governments in many developing countries lack elastic revenue bases, especially with respect to middle and low income citizens. They may also lack the constitutional authority or administrative capacity to levy and collect local taxes. Local services correspondingly tend to be financed by fiscal grants from the centre, whence local governments are no longer self-sufficient, creating a host of problems (such as asymmetric information about local need, 'soft' budget constraints, and dependence of service levels on the vagaries of public finances of the central government) that will be studied in the next Section. An intermediate solution involves local governments financing services by levying user charges, an approach commanding increasing attention in developing countries for infrastructure services. The virtues frequently commended for this approach are that they enhance fiscal autonomy of local governments, thus minimising the problems described above. Less attention has been devoted to the implications for intra-community allocations, to which we now turn.

The key feature of user fee financing (in contrast to local taxes) is their non-coercive character: fees are paid on the basis of voluntary purchase decisions by users. This has two important consequences. First, the government does not need a specialised administration to collect local taxes, limiting deadweight costs. Indeed, we shall assume that these are zero for collection of user fees: this is inessential to the arguments below, which will continue to apply as long as they do not exceed the deadweight costs of central tax revenues.[21]

Second, every citizen has the option of foregoing the service if the fee is excessive, limiting the surplus that local governments can extract from them. Large users can of course still use their political power to evade paying fees for the services they consume. But the voluntary participation constraint for small users restricts the extent of feasible cross-subsidisation. Formally, the optimisation problem faced by the local government in the tax financed regime is subject to the participation constraint:

$$\eta_i \gamma_k v(y_k) - t_k \geq 0, k = s, l. \tag{13}$$

Proposition 3.

(*i*) *Service and fees set by local governments under user-fee-financed decentralisation are as follows:*

$$y_s = y_s^b, t_s = \gamma_s \eta_i v(y_s^b), y_l = \max(y_l^b, \hat{y}_l), t_l = \theta_i(y_l^b - \hat{y}_l) \tag{14}$$

where $\hat{y}_l$ *denotes* $\{\beta_i[\gamma_s \eta_i v(y_s^b) - \theta_i y_s^b]\}/[(1 - \beta_i)\theta_i]$. *Compared with centralisation, service deliveries are larger for both groups. Compared with local-tax-financed decentralisation, service deliveries are higher for small users, while the comparison is ambiguous for large users.*

[21] Note that the efficiency costs of user fees in terms of inducing over or under-use of the service are already incorporated in the analysis below, so the deadweight costs in the user fee mechanism involve only collection costs. These are likely to be much lower than administration and collection of direct taxes, which requires valuation of local properties and monitoring taxable activities of local citizens.

(*ii*) *User-fee-financed decentralisation (weakly) Pareto dominates centralisation: small users are equally well off while large users are better off. It welfare-dominates local-tax-financed decentralisation, i.e., with respect to both efficiency and equity.*

The reasoning is straightforward. Consider the problem of maximising the local government objective function (9) subject to (13) and the budget constraint $\beta_i t_s + (1 - \beta_i) t_l = (1 + \lambda)\theta_i[\beta_i y_s + (1 - \beta_i) y_l]$. The fee t_s for small users will be set at a level which reduces their surplus to zero, while providing them the first-best service level. The financial surplus generated thereby will be used to fund provision of the service to the large users. It pays for a service level $\hat{y}_l$ for large users. This will be the service actually delivered to large users, if it exceeds the first-best level y_l^b. Otherwise the latter will pay the supplemental amount necessary to raise their service to the first-best level.

To prove (*ii*), consider first the welfare comparison with centralisation. It will be simpler to ignore the cost of taxes paid by users in the centralised system; once they are incorporated the welfare of users in that system will become even lower, further reinforcing our conclusion. Small users are exactly as well off, since in either system they receive zero surplus. And large users are better off: this is obvious when $\hat{y}_l \geq y_l^b$, since they receive a larger service and pay nothing. In the other case they receive the first-best service level, the same as in centralisation, and they pay less.[22]

Next consider the comparison with decentralisation financed by local taxes. Here it helps to focus on the case where collection of user fees involves the same deadweight cost λ as local taxes. Small users will get the same (second-best) service level under both systems of financing, since they involve the same burden and allocation of cost. We claim that service provision to large users will either be the same or higher under tax financing. It will obviously be the same in the case where the participation constraint for small users does not bind in the tax financing solution. On the other hand if the participation constraint binds, small users must be paying more under tax financing (since they receive the same service level under both systems), which funds a larger service to large users, as claimed above. Since the latter receive second-best supply or greater under user-fee financing, the service must be over-provided to large users under tax financing. Therefore tax-financing is both less efficient and less equitable. To complete the argument, note that if the collection of user fees involves lower deadweight costs compared to local taxes, the relative performance of the user-fee mechanism improves even further.

A user-fee system administered by a local government subject to local capture thus continues to overprovide the service to the large users at the expense of the small users. But the extent is lower compared with the case of local tax finance. Service levels under the scheme are illustrated in Figure 2. Supply to small users expands uniformly from second-best to first-best because of the reduction in deadweight costs of collection. The same is true for large users in

[22] They pay less than the cost of their service in the decentralised system, being subsidised partly by the small users. Whereas under centralisation bureaucrats earn positive rents from both categories of users, implying that large users pay more than the marginal cost of their service.

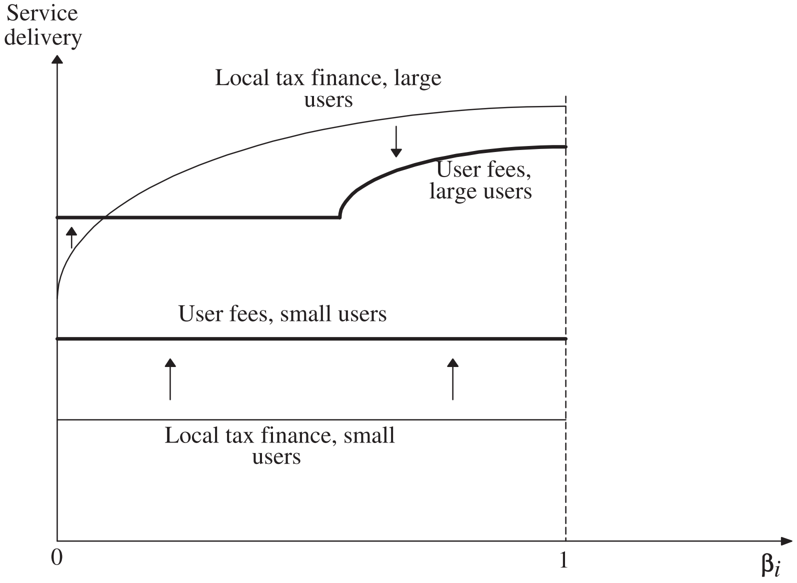

Fig. 2. *Service Delivery Patterns Under User-fee-financed and Local-tax-financed Decentralisation*

regions with negligible (β_i close to 0) small users: supply levels expand from the second-best to the first-best level. In such regions there is an expansion of service deliveries to both groups, compared with the case of tax financing. For regions with higher fraction of small users, there can be over-provision to large users but to a lesser extent under user-fee financing. Hence aggregate service level to communities can shrink as a result of the restriction on the revenue raising capacity of local governments. From a welfare standpoint, however, this is a blessing – it reflects a mitigation of the damaging efficiency and equity effects of local capture.

This explains the welfare ranking of user-fee-financed decentralisation relative to local tax finance or centralisation. The generality of this result is striking: *it holds irrespective of the degree of local capture, the composition of the district, or the realisation of local need and cost shocks.* Of course there are a number of qualifications: it rests on some of the maintained assumptions of the model, such as absence of need to incorporate inter-regional spillovers or redistribution, and adequate capacity of local governments to ensure cost-effective procurement. The argument also utilised the assumption of only two classes of users; we have not checked whether it survives when there are more than two classes. Despite these qualifications, the result illustrates a number of advantages of user fees: lower collection costs, and limited scope for discretionary cross-subsidisation by captured local governments to favour local elites. User fees selected by local governments with purchase decisions subsequently decentralised to individual users permit flexibility of service provision with respect to local cost and need. It is not, however, an optimal mechanism, since it permits some degree of cross-subsidisation and overprovision to large users. This motivates interest in alternative financing mechanisms that restrict discretion of local governments in other ways.

3.3. *Central Grant Financing*

We now consider the third financing mode for local governments: grants from the central government. Suppose that local governments have no revenue raising capabilities at all and receive block grants from the centre to fund infrastructural service allocations. Note that the local government would always prefer a larger grant to a smaller one. This will give rise to an incentive problem between central and local governments: the latter would always like to overstate local need and cost in order to be eligible for a larger grant. Lacking information about local conditions and being unable to monitor service deliveries actually implemented by local governments, the centre will be unable to verify the claims made by local governments. Consequently grants will end up being insensitive to local conditions.

The insensitivity of central grants to local conditions implies that first-best or second-best allocations cannot be implemented, even if all other conditions were ideal (e.g., if local governments were not subject to elite capture). Nevertheless, some flexibility is possible if the grants are not tied to specific categories of services: local governments can then allocate a given budget across different services in response to shifts in relative local needs or costs. Even tied grants admit considerable *de facto* fungibility, allowing them to be spent on alternative services via creative accounting practices. In particular they can be diverted to alternative programmes that happen to be favoured by local elites.

To represent such flexibility in its simplest form, we assume that governments allocate their fiscal resources between the infrastructure service in question, and some alternative services valued by local citizens. The local government can allocate alternative services across the two categories of consumers and we shall assume that the marginal utility of these alternate services are constant and the same for either group (so their value can be replaced by corresponding pecuniary equivalents). The results below will not be qualitatively altered with alternative specifications, e.g., if the alternative services are allocated uniformly across both classes of users, or if marginal utility of users are diminishing with respect to the level of their supply.

It is also useful to clarify that the provision of alternate services did not play any role in the tax financing regime, owing to the assumption of constant marginal cost of local tax finance. In that context, delivery of different services are independent of one another. This is not so under grant financing where the local government allocates a fixed grant budget across different services, so that the true cost of any given service level for citizens is the opportunity cost in terms of alternative public services foregone (rather than additional taxes or user fees paid).

Return now to analyse the outcome of grant financing. Given a fixed (per capita) block grant G, the local government in region i will select an allocation of the given infrastructure service y_s, y_l and pecuniary equivalents S_s, S_l of the value of alternative services for the two classes of users to maximise

$$\beta_i[\gamma_s\eta_i v(y_s) + S_s] + \delta_i^l(1 - \beta_i)[\gamma_l\eta_i v(y_l) + S_l] \tag{15}$$

subject to the budget constraint

$$\beta_i(\theta_i y_s + S_s) + (1 - \beta_i)(\theta_i y_l + S_l) \leq G \tag{16}$$

and the nonnegativity constraints $S_s \geq 0$, $S_l \geq 0$ that arise from the lack of local revenue raising capacity. Given local capture it is immediately evident that small users will receive no alternative services at all: $S_s = 0$. Hence grant income not spent on the infrastructural service will be diverted to the procurement of alternative services that selectively benefit local elites.

Since the budget constraint (16) must bind, it follows that $S_l = (1 - \beta_i)^{-1}\{G - \theta_i[\beta_i y_s + (1 - \beta_i)y_l]\}$: spending more resources on the assigned service means less is available for diversion to the alternative service. *In contrast to the two previous financing modes, therefore, the cost of service delivery at the margin is effectively borne by large rather than small users.* This implies a different pattern of service allocation from the two previous regimes. The problem of the local government reduces to maximisation of

$$\eta_i[\beta_i\gamma_s v(y_s) + \delta_i^l(1 - \beta_i)\gamma_l v(y_l)] + \delta_i^l\{G - \theta_i[\beta_i y_s + (1 - \beta_i)y_l]\} \tag{17}$$

subject to the constraint that $G \geq \theta_i[\beta_i y_s + (1 - \beta_i)y_l]$. If the grant G is large enough, this constraint will not bind, and the service allocations will satisfy the first order condition

$$\gamma_s\eta_i v'(y_s) = \delta_i^l\theta_i; \quad \gamma_l\eta_i v'(y_l) = \theta_i. \tag{18}$$

Large users then get delivered the first-best level, while there is under-provision to small users. This is closer to the pattern under centralisation, rather than the other financing modes of decentralisation.

The implications of a given block grant G on community allocation is described next.

Proposition 4. *The allocation resulting from a block grant G in community i is the following:*

(*i*) *Service delivery for a group k user is* $y_k = f_k Y_i^l$, *where* Y_i^l *is the per capita service level in the community (described further below), and* f_k *is the share of group k, determined as follows:* $f_s = \gamma_s^{-\frac{1}{\alpha}}/[\beta_i\gamma_s^{-\frac{1}{\alpha}} + (1 - \beta_i)\delta_i^l\gamma_l^{-\frac{1}{\alpha}}]$ *and* $\beta_i f_s + (1 - \beta_i)f_l = 1$.

(*ii*) *There exists a threshold need level* η_i^* *that depends on* G, θ_i, δ_i^l *such that the following is true.*[23] *When local need* θ_i *is less than* η_i^*, *the local government is not financially constrained, with the per-capita service level for the community* Y_i^l *equal to the desired level* $Y_i^f(\eta_i, \theta_i)$, *characterised by*

[23] This is given by $\eta_i^*(G, \theta_i, \delta_i^l) = \delta_i^l\theta_i/[L_i v'(G/\theta_i)]$.

$$L_i \eta_i v'(Y_i^f) = \delta_i^l \theta_i \tag{19}$$

where L_i denotes $[\beta_i \gamma_s^{1-\frac{1}{\alpha}} + (1 - \beta_i)\delta_i^l \gamma_l^{1-\frac{1}{\alpha}}]/[\beta_i \gamma_s^{-\frac{1}{\alpha}} + (1 - \beta_i)\delta_i^l \gamma_l^{-\frac{1}{\alpha}}]$. In this case spending on the service is less than the grant G, with the surplus diverted to elite consumption $(S_l > 0)$. When need exceeds η_i^*, the local government is financially constrained, spending it entirely on the service, so $Y_i^l = G/\theta_i < Y_i^f$, and there is no diversion.

The per capita service delivery pattern is

$$Y_i^L(\eta_i, \theta_i, G) = \min[Y_i^f(\eta_i, \theta_i), G/\theta_i]. \tag{20}$$

This restriction in the flexibility of service levels to local conditions in high need states is a distinctive feature of grant-financed decentralisation. It is an outcome of the informational constraints facing central governments while designing fiscal grants, and the incentive of each local government to free-ride off a common revenue pool at the expense of other communities. The severity of these fiscal constraints depends on how large the grant is. We therefore turn to the question of how these grants are determined.

This depends on the political objectives of the central government, and the way that they fund the grants. The central government may also be subject to capture by elites, to an extent that may bear no obvious relation to the extent of local capture, as argued in Bardhan and Mookherjee (1999, 2000*a*). So letting δ^c denote the degree of capture at the central level, the objective of the central government is

$$\Sigma_i N_i[\beta_i U_{si} + \delta^c(1 - \beta_i)U_{li}]. \tag{21}$$

As for financing patterns, it is well known that for a variety of reasons, both including political will and administrative ease, most developing countries rely primarily on indirect (sales, excise and customs duties) rather than direct taxes.[24] Owing to their regressive nature, we shall assume that small users bear a burden that is proportionately greater or the same as the burden borne by large users. Let $1 - \psi \in (0, 1)$ denote the asymmetry in tax burden, i.e., if $\psi = 0$ the burden falls exclusively on small users, whereas it is shared evenly if $\psi = 1$. Then the objective of the central government as a function of the grant allocation $G_1, G_2, \ldots$ to different communities reduces to

$$\begin{aligned} V^c(G_1, G_2, \ldots) \equiv \Sigma_i N_i \mathcal{E}_{\eta_i,\theta_i}\{&\beta_i \gamma_s \eta_i v(f_s Y_i^L) \\ &+ \delta^c[(1 - \beta_i)\gamma_l \eta_i v(f_l Y_i^L) + G_i - \theta_i Y_i^L] \\ &- [1 + \psi(\delta^c - 1)(1 - \beta_i)](1 + \lambda)G_i\}. \end{aligned}$$

This can be expressed as the sum of separate objective functions for different regions:

$$V^c(G_1, G_2, \ldots) \equiv \Sigma_i N_i V_i^c(G_i) \tag{22}$$

[24] See the evidence cited in Ahmad and Stern (1984), Newbery and Stern (1987), Das-Gupta and Mookherjee (1998).

where the objective function corresponding to community i is a function of the grant to that community alone:

$$\begin{aligned} V_i^c(G_i) \equiv \mathcal{E}_{\eta_i,\theta_i}\{&\beta_i\gamma_s\eta_i v(f_s Y_i^L) + \delta^c[(1-\beta_i)\gamma_l\eta_i v(f_l Y_i^L) \\ &+ G_i - \theta_i Y_i^L] - [1+\psi(\delta^c-1)(1-\beta_i)](1+\lambda)G_i\}\end{aligned}. \tag{23}$$

The community grant G_i will be selected to maximise (23). In making this decision, the central government incorporates its expectations of how local governments will allocate any given grant level within their respective communities.

The analysis of optimal community grants is somewhat complicated, and so we omit some of the technical steps (which are available in the working paper version of this article Bardhan and Mookherjee (2000*c*)). The overall implications are summarised below.

PROPOSITION 5. *With decentralisation financed entirely by central grants:*

(*a*) *Region i will be financially constrained with positive probability if*

$$(1+\lambda) > [(\delta^c)^{-1} + \psi(1-\beta_i)(1-(\delta^c)^{-1})]^{-1}. \tag{24}$$

In this case, region i will be financially constrained if and only if local need shock η_i exceeds the threshold η_i^*.

(*b*) *In low need states where region i is not financially constrained, large users are provided first-best service levels (besides the benefits of diverted funds), while small users are underprovided relative to the first-best to an extent depending on local capture. In financially constrained states, service levels are the same as at the threshold state η_i^*, and no funds are diverted.*

(*c*) *Service delivery levels for either group are smaller in all states compared with user-fee-financed decentralisation.*

(*d*) *If the deadweight cost of taxes λ is sufficiently large, grant-financed decentralisation is less efficient compared with either centralisation, or decentralisation financed by central taxes or by user fees. If λ is sufficiently small, local and central capture $(\delta_i^l - 1), \psi(\delta^c - 1)$ sufficiently close to zero, then grant-financed decentralisation approaches the first-best.*

Note the importance of financing constraints faced by the central government, represented by λ. Even with perfectly accountable governments, financing constraints at the central level will lead to service underprovision with grant financing, unlike decentralisation based on user fees. As λ rises, service levels will progressively shrink as central grants dry up. For λ sufficiently large, service levels will decline precipitously, causing performance to drop below centralisation as well. At the other extreme, if collection at the centre is efficient and λ is close to zero, and governments are sufficiently accountable at both levels, the outcome of grant-financed decentralisation will approach the first-best.

In particular, note that grant financing may be dominated by user-fee financing under appropriate conditions (e.g., λ sufficiently large), while under others grant financing may be more or almost as efficient than user fees. To gain further insight into the relevant trade-offs, we compare the resulting patterns of service deliveries with centralisation and user-fee-financed decentralisation. These are depicted in Figures 3 to 6. Figure 3 compares deliveries with those under centralisation for regions where $\beta_i < \beta_1^*$, where the threshold β_1^* is defined by the condition $\delta_i^l = [1 - (1 - \beta_1^*)(\gamma_l/\gamma_s - 1)\beta_1^{*-1}]^{-1}$, i.e., service undeprovision to small users is the same as in centralisation. In low β_i regions, grant financing expands supplies to small users in low need states where the local government is not financially constrained, while supplies to large users is unaffected. But in high-need states where local financing constraints bind, service levels may shrink for both groups as a result of transition to grant-financed decentralisation. The overall effect on service levels and efficiency thus depends on the severity of the local financing constraints, as explained above. One apparent benefit of decentralisation in these regions is that it improves equity in service levels (in low need states). This may however not be mirrored in a genuine improvement in equity since small users may ultimately bear a greater financial burden under decentralisation (e.g., if they bear a disproportionate share of the burden of central taxes).

On the other hand in regions where the fraction of small users β_i is larger (depicted in Figure 4), service allocations to small users shrink under decentralisation even when local governments are financially constrained, resulting in a less equitable outcome. In this case, service levels shrink for both classes of users (except large users in low need states, who receive the same service). Here grant-financed decentralisation hurts growth, efficiency as well as equity. Indeed, it may be Pareto-inferior to centralisation: large users may be worse off

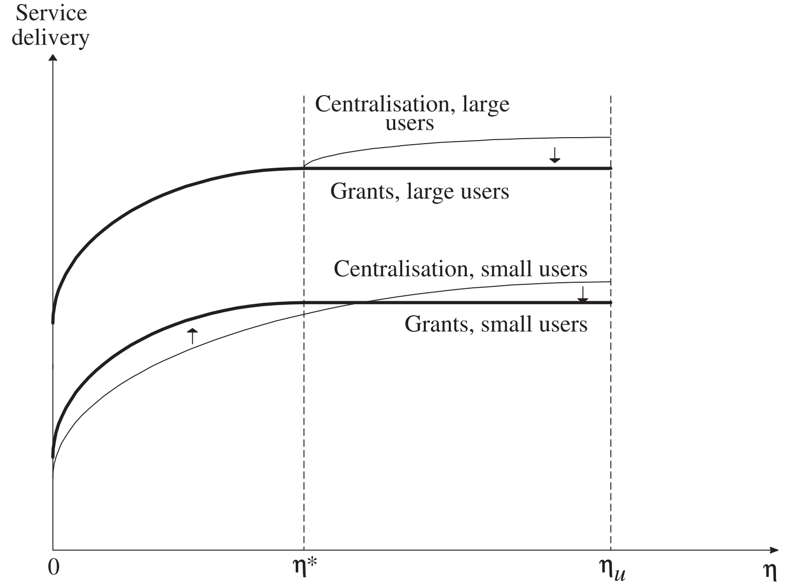

Fig. 3. *Service Patterns Under Centralisation and Grant-financed Decentralisation in Regions With* $\beta_i < \beta_1^*$

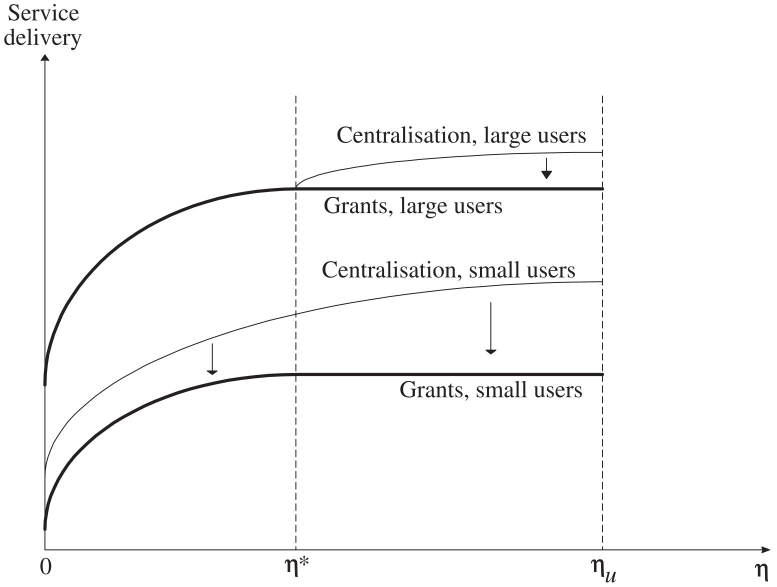

Fig. 4. *Service Patterns Under Centralisation and Grant-financed Decentralisation in Regions With* $\beta_i > \beta_1^*$

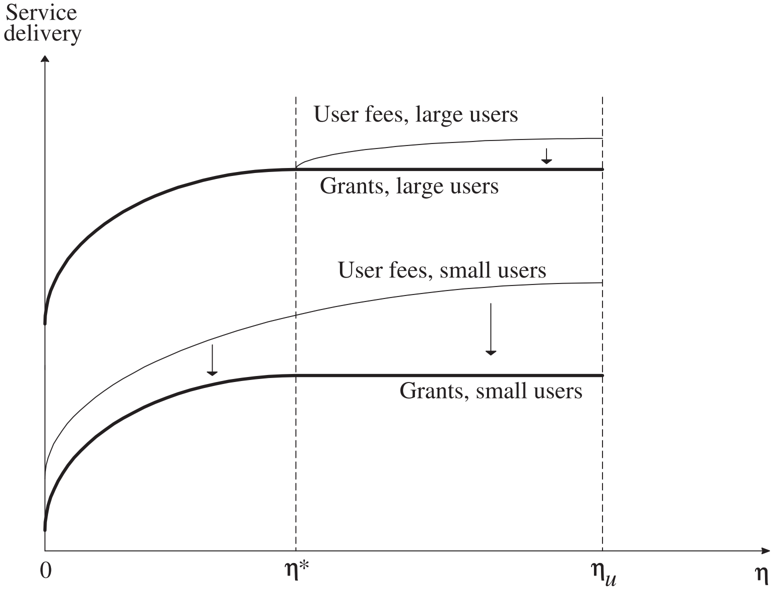

Fig. 5. *Service Patterns Under User-fee-financed and Grant-financed Decentralisation in Regions with* $\beta_i < \beta_1^*$

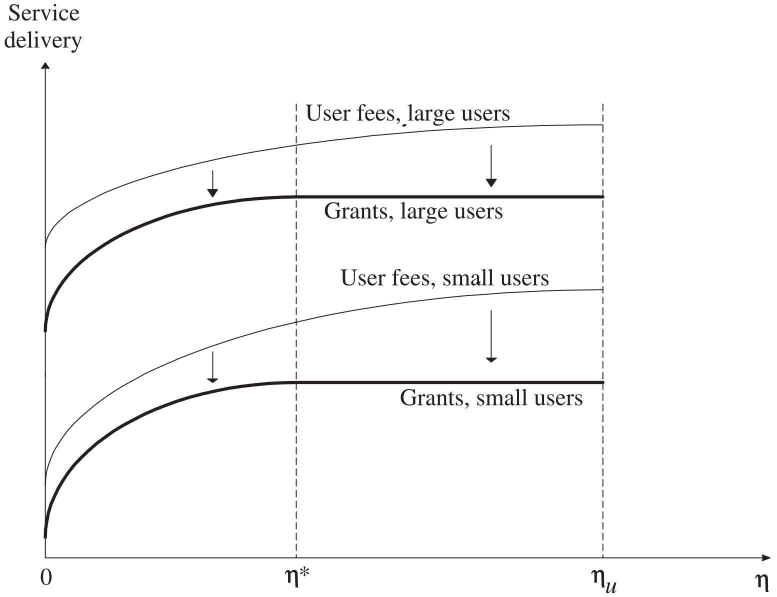

Fig. 6. *Service Patterns Under User-fee-financed and Grant-financed Decentralisation in Regions with* $\beta_i > \beta_2^*$

even if they bear a negligible fraction of the burden of central taxes, as a consequence of shrinking central grants that dry up service deliveries. This is a case where transition to decentralisation will cause both categories of users to regret the absence of the corruption which 'lubricated' the centralised system in the past!

We turn finally to the comparison with user-fee financed decentralisation, which we have shown above dominated local-tax financing. Note that user-fee financing generates efficient provision of the service if and only if β_i is small enough, whence the burden of financing over-provision on small users tends to be excessive. For instance in the case where α, the elasticity of marginal utility of consumption, lies between 0 and -1, the relevant threshold β_2^* is defined by the condition $\beta_2^*[(1+\alpha)^{-1}-1]/(1-\beta_2^*) = (\gamma_s/\gamma_l)^{\frac{1}{\alpha}}$. For β_i smaller than this threshold, both categories of users are served at the first-best level under user financing, and small users bear the entire fiscal burden. For β_i higher, large users are overprovided, while small users continue to be efficiently funded. See Figures 5 and 6 for these two cases respectively. It therefore follows that in all cases, grant financing shrinks service levels to both categories of users relative to user-fee financing, with the exception of large users in low β_i regions when their local governments are not financially constrained. Hence *the effect of not devolving revenue raising powers to local goverments in step with their expenditure responsibilities causes an unambiguous reduction in the level of services in all regions, irrespective of patterns of political accountability.* This is again consistent with the empirical results of Estache and Sinha (1995). The reduction in service levels may however be efficiency enhancing, as they constrain the tendency for large users to be overprovided under user-fee financing. Since such overprovision is paid for by the small users, it may improve local equity as well. The problem with grant financing on the other hand is the tendency for small users to be under-served if local governments are susceptible to capture, or both categories to get under-served in high-need states when financing constraints bind. The severity of the latter problems depend respectively on the extent of local capture, and on the costs of raising central taxes. If they are not very acute, grant financing may conceivably end up dominating user-fee financing.

4. Related Literature

As mentioned in the Introduction, the issue of corruption and lack of accountability has largely been ignored by previous literature on fiscal decentralisation, with the exception of Seabright (1996) and Tommasi and Weinschelbaum (1999). These authors focus on lack of accountability as the principal drawback of centralisation, which has to be traded off against interjurisdictional coordination problems inherent in decentralisation. In contrast, our theory is based on the view that centralised systems may be more or less accountable than local governments, depending on the nature of political institutions. The basis for this view has been argued by Bardhan and Mookherjee (1999, 2000*a*). Concern has frequently been expressed in many developing countries regarding the possible danger of worsened intracommunity allocations under decentralisation owing to capture of local governments. At the same time our model focuses less on problems of interjurisdictional coordination.[25]

Our result concerning the value of constraining financing options of local governments bears some resemblance to the arguments of Brennan and Buchanan (1980) concerning the need to impose constitutional constraints on Leviathan-like governments. Yet there are many significant differences between our respective approaches and results. First, Brennan and Buchanan adopt the Leviathan assumption universally, wherein all governments are depicted as seeking to maximise surplus of revenues over public good supply for their own benefit. In contrast we derive objectives of decision makers from underlying information and control structures in either system. In our model Leviathan is an apt characterisation of bureaucrats in the centralised regime, but not of local governments in the decentralised system. The difference arises from the role of political competition in the latter. The main argument in favour of decentralisation is precisely that it

[25] For instance there are no capacity constraints on service delivery across different communities. We also abstract from the possibility that local governments may possess less administrative or technical competence relative to central bureaucrats. Our model does however accommodate economies of scale in service provision across communities, which motivates production to be concentrated in a single utility from which local governments procure the service. It also allows for fiscal externalities across jurisdictions, i.e., the tendency for local communities to free-ride off revenues raised by the central government in the grant-financed system.

limits the Leviathan tendency of centralised bureaucrats.[26] The problem with local governments is not their Leviathanism but favouritism with regard to one category of citizens. Moreover, our result concerning the superiority of decentralisation under appropriate restrictions on financing authority of local governments, does not find a parallel in their work. Brennan and Buchanan lay greater emphasis instead on the need to constrain financing options of governments at the central level rather than at the local level, owing to greater competitive discipline on the Leviathanite tendencies allowed by citizen mobility across jurisdictions.[27]

5. Concluding Comments

This article has studied the tradeoffs between allocation distortions resulting from monopoly power of unregulated and corrupt bureaucrats in a centralised delivery system, and the tendency for local governments to be captured by local elites under decentralisation. The key point is that the effects of decentralising service delivery will depend on the method chosen for financing local governments. Existing empirical results suggest that expenditure decentralisation not accompanied by revenue decentralisation limit the expansionary effect of decentralisation on service levels. Our model provides an explanation for this pattern, and at the same time urges caution in inferring that greater revenue decentralisation would be welfare enhancing. Local capture tends to be manifested in service overprovision to local elites, at the expense of elites, which is both inefficient and inequitable. Accordingly restraints on the revenue-raising capability of local governments can limit the extent of such resource misallocations.

User-fee-financing mechanisms are particularly notable in this connection: the voluntariness of such mechanisms in contrast to the coercive character of local taxes limit the extent of regressive redistributions that elites can employ in their favour. In our model user fees ensure that decentralisation generates higher efficiency and equity compared to centralisation, irrespective of the extent of local capture. Compared with the more traditional form of financing, i.e., inter-governmental fiscal grants, user fees have the added advantage of enhancing fiscal autonomy of local governments. This enables service allocations to be sensitive to random fluctuations in local costs and needs, particularly when such flexibility is most useful (i.e., when local need is high). They also ensure higher service deliveries compared with grant financing, owing to the avoidance of asymmetric information, inter-community free-riding and bargaining distortions inherent in a

[26] In contrast, the arguments made by Brennan and Buchanan in favour of fiscal decentralisation stem from the Tiebout-like competitive pressures operating on local governments when citizens can move costlessly from one jurisdiction to another (a phenomenon we abstract from on the grounds of limited relevance, especially in developing countries). Moreover, Brennan and Buchanan assert the tradeoffs between such gains in government accountability and associated costs of lack of scale economies and existence of interjurisdictional externalities. In our context the principal tradeoff is with the tendency for local governments to be captured by special interest groups.

[27] The relevance of citizen mobility as a source of inducing greater accountability of governments is questionable especially in the context of developing countries, where mobility costs are high, services are scarce, and recent migrants to communities face great difficulty in securing access to local public services.

system of intergovernmental fiscal grants. Of course user fee mechanisms have a number of shortcomings which our model abstracted from: for instance when redistribution across communities is an important objective, or if a significant proportion of intended beneficiaries do not have the means to pay for the service.

Apart from the normative results, our model also provides a number of detailed predictions concerning the impact of decentralisation on service allocations and their financing, which are empirically testable. We hope that future empirical analyses of fiscal decentralisation in developing countries will be carried out to test these predictions.

Our model abstracted from problems of interregional spillovers of decisions made by local governments, and possible lack of expertise at local levels. Both of these may be important in practice, and need to be evaluated independently in assessing the effects of decentralisation. Spillovers might naturally arise in the areas of roads, telecommunications, schools and public health. Even in the context of water resources, spillovers will arise if there are aggregate capacity constraints that bind. In all these cases, decentralisation will require coordination of decisions made independently by different local governments, involving either central interventions, establishment of resource sharing formulae, or market-like mechanisms. Lack of managerial and technical expertise at the local level may prevent cost-effective provision of the service within regions. For instance, if local government officials are not informed about the realisation of marginal costs of serving their community, managers of the central production enterprise may be able to earn rents by exploiting their specialised information, resulting in additional distortions under decentralisation. Cost-effective procurement may also be vitiated if local government officials do not have much bargaining power when dealing with service providers, allowing the latter to earn monopoly rents. In the presence of either of these problems, the performance of decentralised regimes will deteriorate further.

We considered three polar modes of financing most commonly employed in developing countries (Dillinger, 1995). Mixed modes of financing may also be worth exploring in this context, e.g., where local governments rely on a mixture of local user charges and central grants, which might dominate either polar mode. We also restricted attention to unrestricted fiscal grants, which allowed some degree of flexibility in local service delivery, at the cost of allowing diversions of surplus resources to local elites in low-need states. Grants tied to expenditures on specific services restrict both flexibility in service delivery and scope for diversion of unspent funds to other less important social purposes. An intermediate form of grant finance involves matching grants tied to specific services, which combine advantages of providing some degree of flexibility in service delivery, while limiting scope for diversion. Clearly the welfare implications of a richer set of financing options than analysed in this article deserve to be explored in future research. Finally, it may also be worthwhile to explore additional variants of decentralisation such as privatisation of the service delivery process (where private delivery companies are subject to a combination of central and local government regulation), or yardstick competition between different local governments.

University of California, Berkeley
Boston University

Submitted: 26 February 2002
Accepted: 16 November 2004

References

Ahluwalia, M. (1998). 'Infrastructure development in India's reforms', in (I.M.D. Little ed.) *Indian Economic Reforms and Development: Essays for Manmohan Singh*, pp. 87–121, Delhi, Oxford and New York: Oxford University Press.

Ahmad, E. and Stern, N. (1984). 'The theory of tax reform and Indian indirect taxes', *Journal of Public Economics*, vol. 25(3), pp. 259–98.

Banerjee, A. (1997). 'A theory of misgovernance', *Quarterly Journal of Economics*, vol. 62, pp. 1289–332.

Bardhan, P. (1996). 'Efficiency, equity and poverty alleviation: policy issues in less developed countries', ECONOMIC JOURNAL, vol. 106 (September), pp. 1344–56.

Bardhan, P. (2002). 'Decentralization of governance and development', *Journal of Economic Perspectives*, vol. 16(4), pp. 185–206.

Bardhan, P. and Mookherjee, D. (1999). 'Relative capture of local and central governments: an essay in the political economy of decentralization', Working Paper, Institute for Economic Development, Boston University.

Bardhan, P. and Mookherjee, D. (2000*a*). 'Capture and governance at local and national levels', *American Economic Review*, vol. 90(2), (May), pp. 135–9.

Bardhan, P. and Mookherjee, D. (2000*b*). 'Decentralizing anti-poverty program delivery in developing countries', Working Paper, Institute for Economic Development, Boston University, forthcoming, *Journal of Public Economics.*

Bardhan, P. and Mookherjee, D. (2000*c*). 'Corruption and decentralization of infrastructure delivery in developing countries', Working Paper, Institute for Economic Development, Boston University.

Bardhan, P. and Mookherjee, D. (eds) (2005). *Decentralization to Local Governments in Developing Countries: A Comparative Perspective*, forthcoming, Cambridge, MA.: MIT Press.

Baron, D. and Myerson, R. (1982). 'Regulating a monopolist with unknown cost', *Econometrica*, vol. 50(4), (July), pp. 911–30.

Baron, D. (1994). 'Electoral competition with informed and uninformed voters', *American Political Science Review*, vol. 88, pp. 33–47.

Besley, T. (1989). 'Targeting taxes and transfers: administrative costs and policy design in developing economies', Development Studies Working Paper 146, Princeton University, Woodrow Wilson School.

Besley, T. and Kanbur, R. (1993). 'Principles of targeting', in (M. Lipton and J. ven der Gaag, eds.) *Including the Poor*, Washington DC: World Bank.

Besley, T. and Coate, S. (2003). 'Centralized versus decentralized provision of local public goods: a political economy analysis', *Journal of Public Economics*, vol. 87(12), pp. 2611–37.

Bird, R. (1995). 'Decentralizing infrastructure: for good or ill?', in (A. Estache ed.) (1995), Washington, DC: World Bank.

Bolton, P. and Roland, R. (1997). 'The breakup of nations: a political economy analysis', *Quarterly Journal of Economics*, vol. 62, pp. 1057–90.

Brennan, G. and Buchanan, J. (1980). *The Power to Tax: Analytical Foundations of a Fiscal Constitution*, Cambridge: Cambridge University Press.

Cremer, J., Estache, A. and Seabright, P. (1995). 'The decentralisation of public services: lessons from the theory of the firm', in (A. Estache, ed.) (1995), pp. 98–118.

Crook, R. and Manor, J. (1998). *Democracy and Decentralisation in South Asia and West Africa*, Cambridge: Cambridge University Press.

Das-Gupta, A. and Mookherjee, D. (1998). *Incentive and Institutional Reforms in Tax Enforcement: An Analysis of Developing Country Experience*, New Delhi: Oxford University Press.

Dillinger, B. (1995). 'Decentralization, politics and public service', in (A. Estache ed.) (1995), pp. 5–21.

Dreze, J. and Saran, M. (1995). 'Primary education and economic development in China and India: overview and two case studies', in (K. Basu, P. Pattanaik and K. Suzumura), *Choice, Welfare, and Development: A Festchrift in Honor of Amartya K. Sen*, Oxford: Clarendon Press.

Estache, A. (ed.) (1995). *Decentralizing Infrastructure: Advantages and Limitations*, Washington D.C.: World Bank Discussion Paper no. 290.

Estache, A. and Sinha, S. (1995). 'Does decentralisation increase public infrastructure expenditure?', in (A. Estache ed.) (1995), pp. 63–79.

Fisman, R. and Gatti, R. (2002). 'Decentralization and corruption: evidence across countries', *Journal of Public Economics*, vol. 83, pp. 325–45.

Grosh, M.E. (1991). 'The household survey as a tool for policy change: lessons from the Jamaican survey of living conditions', Washington D.C.: LSMS Working Paper 80, World Bank.

Grosh, M.E. (1995). 'Towards quantifying the tradeoff: administrative costs and targeting accuracy', in (D. Van de Walle and K. Nead, eds.), *Public Spending and the Poor: Theory and Evidence*, pp. 450–88, Baltimore: Johns Hopkins University Press.

Grossman, G. and Helpman, E. (1996). 'Electoral competition and special interest politics', *Review of Economic Studies*, vol. 63, pp. 265–86.

Inman, R. and Rubinfeld, D. (1996). 'Designing tax policies in federalist economies: an overview', *Journal of Public Economics*, vol. 60, pp. 307–34.

Inman, R. and Rubinfeld, D. (1997). 'Rethinking federalism', *Journal of Economic Perspectives*, vol. 11(4), pp. 43–64.

Laffont, J. and Pouyet, J. (2000). 'The subsidiarity bias in regulation', Working Paper, University of Toulouse.

Laffont, J. and Tirole, J. (1993). *A Theory of Incentives in Procurement and Regulation*, Cambridge, MA: MIT Press.

Lieten, G. (1996). 'Panchayats in western Uttar Pradesh', *Economic and Political Weekly*, September 28, pp. 2700–5.

Lipton, M. and Ravallion, M. (1995). 'Poverty and policy', in (J. Behrman and T.N. Srinivasan eds.), *Handbook of Development Economics*, vol. 3, Chapter 41, Amsterdam: North-Holland.

Litvack, J., Ahmed, J. and Bird, R. (1998). 'Rethinking decentralization at the World Bank', Washington D.C.: Discussion Paper, World Bank.

Lockwood, B. (1998). 'Distributive politics and the benefits of decentralisation', CSGR Working Paper No. 10/98.

Manor, J. (1999). *The Political Economy of Democratic Decentralisation*, Washington DC: The World Bank.

Mathew, G. and Nayak, R. (1996). 'Panchayats at work: what it means for the oppressed?', *Economic and Political Weekly*, July 6, pp. 1765–71.

Musgrave, R. and Musgrave, P. (1984). *Public Finance in Theory and Practice*, 4th Edn, New York: McGraw Hill.

Newbery, D. and Stern, N. (1987). *The Theory of Taxation for Developing Countries*, World Bank: Oxford University Press.

Oates, W. (1972). *Fiscal Federalism*, New York: Harcourt, Brace and Jovanovich.

Prud'homme, P. (1995). 'The dangers of decentralization', *World Bank Research Observer*, vol. 10, pp. 201–20.

Seabright, P. (1996). 'Accountability and decentralisation in government: an incomplete contracts model', *European Economic Review*, vol. 40 (1), pp. 61–89.

Tanzi, V. (1996). 'Fiscal federalism and efficiency: a review of some efficiency and macroeconomic aspects', in (M. Bruno and B. Pleskovic eds.), *Annual World Bank Conference on Development Economics 1996*, Washington DC: The World Bank.

Tommasi, M. and Weinschelbaum, F. (1999). 'A principal-agent building block for the study of decentralisation and integration'. Working Paper, Universidad de San Andres, Buenos Aires.

Van de Walle, D. and Nead, K., (eds.) (1995). *Public Spending and the Poor: Theory and Evidence*, Baltimore: Johns Hopkins University Press.

Wade, R. (1985). 'The market for public office: why the Indian state is not better at development', *World Development*, vol. 13(4), pp. 467–97.

Wade, R. (1997). 'How infrastructure agencies motivate staff: canal irrigation in India and the Republic of Korea', in (Ashoka Mody, ed.), *Infrastructure Strategies in East Asia*, Washington D.C.: World Bank.

World Development Report, (1990, 1994, 1997, 2004). Washington DC: World Bank and Oxford University Press.