THE ECONOMICS OF FISCAL DECENTRALIZATION

Duc Hong Vo

UWA Business School, University of Western Australia

Abstract. There is no complete overview or discussion of the literature of the economics of federalism and fiscal decentralization, even though scholarly interest in the topic has been increasing significantly over recent years. This paper provides a general, brief but comprehensive overview of the main insights from the literature on fiscal federalism and decentralization. In doing so, literature on fiscal federalism and decentralization is grouped into two main approaches: ‘first generation approach’ and ‘an emerging second generation approach’. The discussion generally covers the two notions of fiscal decentralization: ‘fiscal autonomy’ and ‘fiscal importance’ of subnational governments as the background of the most recently developed index of fiscal decentralization in Vo. The relevance of this discussion to any further development of a fiscal decentralization index is briefly noted.

Keywords. Decentralization theorem; Musgravian fiscal framework; Public choice; Tiebout sorting

Decentralisation has, not only an administrative value, but also a civic dimension, since it increases the opportunities for citizens to take interest in public affairs; it makes them get accustomed to using freedom. And from the accumulation of these local, active, persnickety freedoms, is born the most efficient counterweight against the claims of the central government, even if it were supported by an impersonal, collective will.


1. Introduction

The term ‘fiscal decentralization’ refers to the devolution of authority for public finances and the delivery of government services from the national to subnational levels (Tanzi, 1995). This devolution is related to the four main interrelationships among levels of government regarding fiscal issues, namely the responsibility for (i) expenditure decisions; (ii) taxing and revenue-raising powers; (iii) subnational borrowings and (iv) intergovernmental fiscal transfers. Fiscal decentralization is
currently an issue of considerable practical importance facing many developing economies and has been advocated by international economic bodies, such as the World Bank and the Organisation for Economic Cooperation and Development (World Bank, 2003). Moreover, the academic study of fiscal decentralization is a field of vigorous research activity and a number of attempts have been made to systematically understand the key economic principles of this area (see, for example, McLure, 1998; Collins, 2001; Boadway, 2003; Bird, 2004). The number of publications on fiscal federalism is much smaller than those for the other key topics such as unemployment, inflation, exchange rate, interest rate and foreign direct investment. However, the area is growing rapidly. In fact, with an average growth rate of about 28% per annum, fiscal federalism is the second-fastest growing field among these important fields.

The topic of fiscal federalism was generally introduced into public finance theory in the mid twentieth century, which opened the door to the systematic analysis of fiscal decentralization. The main normative question associated with this subject concerns the extent to which fiscal powers and responsibilities should be devolved from higher to lower levels of government. The level of analysis associated with this question has now developed to the extent where scholars have started to distinguish between first and second generation theories of fiscal decentralization (Oates, 2005, p. 349). The second generation theory of fiscal decentralization is only newly emerging and it does not yet represent a coherent system of analysis, or at least, it has not been well established.

It is noted that the focus of this paper is exclusively on theoretical developments on the issue of fiscal federalism and decentralization, rather than empirical developments. In this paper, the first generation approach to public economics of decentralization is overviewed in Section 2, with attention given to the seminal contributions by distinguished scholars such as Richard Musgrave, Wallace Oates, Charles Tiebout and James Buchanan. The treatment of the four main issues, namely the assignment of responsibility for service provision, the assignment of taxing powers, intergovernmental fiscal transfers and subnational borrowing, in the first generation theory of fiscal decentralization is discussed in Section 3. A brief discussion of the newly emerging second generation approaches to fiscal decentralization, which draws on ideas from outside the public finance literature, is included in Section 4, which is followed by conclusions in Section 5.

2. First Generation Approaches to the Public Economics of Decentralization

A policy of fiscal decentralization is directed towards the transfer of fiscal powers and responsibilities from the national to subnational governments (SNGs). While fiscal centralization is often a response to the demands of national unity, fiscal decentralization may be seen as a response to the demands for diversity and accountability within the community. On balance, the first generation theorists investigating fiscal federalism tended to associate the process of fiscal decentralization with an enhancement in the overall degree of public sector responsiveness to a public demand and, ultimately, to an improvement in the
economic efficiency of public economic activities by better linking resource allocation with public preferences.

Among many different economic ideas about decentralization of public functions to SNGs, and the associated issue of public finances under decentralized systems, seminal contributions were made by Tiebout (1956), Musgrave (1959) and Oates (1972), all of whom laid the strong foundation for significant discussions of fiscal decentralization. Olson (1969), through his concept of fiscal equivalence, also made an important contribution. These studies, in conjunction with the public choice approach to multi-tier government initially developed by Brennan and Buchanan (1980) in *The Power to Tax – Analytical Foundations of a Fiscal Constitution*, represent seminal works in the first generation literature on fiscal decentralization.

2.1 Musgravian Fiscal Framework and Federalism

In the influential book *The Theory of Public Finance – A Study in Public Economy*, Musgrave (1959) introduced three different branches or categories of public finance: economic stabilization, income distribution and resources allocation. These branches have come to represent the benchmarks from which issues in public economics are treated by the non-public choice group within their first generation studies of fiscal decentralization. In general terms, each branch is individually subject to consistent theoretical analysis. Between these branches, analytical consistency is more difficult because of diverse and subjective assessments of the relative importance of stabilization, income distributions and efficiency. In the specific public finance perspective on federalism, the Musgravian branches of public finance proved useful in setting the constraints to fiscal decentralization (stabilization and distribution) and the potential benefits of fiscal decentralization (efficiency).

The first branch is economic stabilization. Fiscal and monetary policies may contribute to stable economic development. The key issue for theory is whether fiscal decentralization enhances or detracts from economic stability and compromises the general macroeconomic goals of governments. The received view among first generation theorists is clear: decentralization of fiscal arrangement does not serve to enhance the macro stabilization objective; rather, macro stabilization represents a constraint on the feasible degree to which fiscal powers can be devolved to SNGs (Oates, 1972, p. 7).

The second branch of public economics concerns income distribution. Economic goods are produced and governments typically play some role in modifying the market determined distribution of goods across members of a given society. Taxes may be progressive and welfare services may target lower income members of society. Again, it appears that decentralization of fiscal arrangements does not tend to systematically advance society’s income distribution objectives because economic interdependences between the economies of subnational jurisdictions act to reduce any diversity in the distribution of goods that would exist across lower level governments. In the absence of diversity in redistribution goals between SNGs, if the demand for redistribution is nationally uniform, the case
for assigning responsibility for redistribution to SNGs is weakened. In short, an income redistribution policy has a greater chance of success if it is carried out at the national level (Oates, 1972, p. 9).

It is in the final branch of public economics – resources allocation – where the importance of fiscal decentralization emerges in the work of first generation theorists. This is so for two main reasons. First, scarce resources should be more efficiently allocated under a decentralized fiscal system, as SNGs may be in a better position of understanding how to maximize benefits from the use of resources in their localities. Second, the character of ‘impure’ or local public goods adds a local congestion dimension to service provision that national governments may not be well placed to manage. In short, if the ‘local’ dimension of public goods is ignored, provision of such goods at the same level across all regions will compromise efficiency when preferences for local public goods differ by regions, jurisdictions or levels of government. As a result, SNGs have the potential to play a major role in efficient resource allocation. In this branch of public economics, the intention to mobilize resources regionally is recognized so that the preferences and tastes of heterogeneous inhabitants can be better served.

Consequently, within the Musgravian framework, the first generation theory of fiscal decentralization primarily emphasizes the brand of public economics concerned with the efficiency of resource allocation. When the first generation theory considers fiscal decentralization, such as in the assignment of expenditure responsibility and revenue-raising powers, it is primarily aimed at achieving efficient resource allocation which maximizes welfares for constituencies. In addition, intergovernmental fiscal transfer systems from the national government to SNGs and the autonomy of SNGs’ borrowings are also discussed in a similar manner in the first generation theory of fiscal decentralization. While the branches of public finance concerned with economic stabilization (to maintain an economy at high levels of employment and stable price level) and income distribution (to achieve horizontal equity and vertical equity in a manner that avoids inefficiency, and to avoid mobility of households and firms across jurisdictions) are considered within the first generation theory of fiscal decentralization, they are generally considered as secondary to the degree of fiscal decentralization.

2.2 Tiebout Sorting

Samuelson’s optimal level of public goods provision cannot be obtained since there is no mechanism for preference revelation. Tiebout (1956) introduced the notion of ‘impure’ or local public goods to the theory of public finance. He did so to analyse political and fiscal decentralization in terms of competition among localities, with the mobility of citizens between localities providing the mechanism for preference revelation. The essence of the Tiebout hypothesis is that consumer demand for local public goods can be revealed when citizens choose the jurisdiction which provides them with the best net benefit. The hypothesis states that, with mobility, consumer-voter’s preferences can be revealed and consumers will end up at, or at least close to, the point where their demand for impure public goods is met with
due recognition of the costs of supplying this demand. This is now generally known as ‘Tiebout sorting’. In particular, this notion of Tiebout’s was primarily a response to the problem of the under-provision of public goods as posed by Paul Samuelson. Tiebout shows that when public goods are provided by competing regions, sorting according to preferences will induce an efficient provision.

Tiebout’s notion of ‘impure’ public goods differs from global or pure public goods, as developed by Samuelson (1954), which are defined by two main characteristics: ‘non-excludability’ and ‘non-rivalry’ in consumption. Non-excludability means that once public goods are produced, no one can be excluded from consuming them. Non-rivalry requires that the marginal benefits that any consumer derives from the consumption of public goods do not change if the number of consumers increases. National defence is the classic example of the pure public good. Once this service is produced, it is impossible, or at least inefficient, to exclude any person in a community from its consumption (Buchanan, 1999). Additional residents are added to a jurisdiction and consume pure public goods without any additional costs. The characteristic of non-rivalry of pure public goods has implied a consequence which was mentioned by Samuelson (1954).

In response to Samuelson’s concerns over preference revelation of public goods, Tiebout (1956) argued that many public goods, such as parks, are actually ‘impure’ because they are subject to congestion (i.e. they are not rival, or progressively less rival, beyond some locally defined boundary). When the number of users increases, public goods consumption may become congested.

Tiebout’s vision is in contrast with Samuelson’s work, which emphasizes that preferences for publicly provided goods are not revealed because the revelation would result in an increase in payment to the citizen who declares his preference without any increase in the quantity of the public good provided (Hamilton, 1987). In short, the Samuelsonian approach considers ‘pure’ public goods whereas the Tiebout approach emphasizes the ‘impurity’ of public goods. Mobility among jurisdictions is possible even though there are some typical obstacles to consumer mobility, such as job commitment and family connection (Oates, 2005, p. 350). In a decentralized society, consumers will effectively sort themselves into homogeneous groups whose demands for certain public goods and services are the same (Oates, 2005, p. 353). As such, decentralized provision of local public goods and services is essential.

A significant aspect of Tiebout sorting is its relevance to policymakers, and the consequent implications for fiscal arrangements. Policies to encourage residential mobility and develop knowledge among residents of public benefits and costs will encourage migration which acts as a force that contributes to improvements in the efficiency of government expenditures. It is also argued that the higher the revealed degree of mobility of households from region to region, the higher the efficiency of the allocation of resources, other things held constant (Tiebout, 1956). This is because there would be no mobility of households and firms when all subnational units are in an equilibrium state – benefits provided would be consistent with costs. As a result, due to the pressure on governments from this mobility, production efficiency will be enhanced.
In summary, Tiebout has made two main contributions to the study of fiscal decentralization. First, he introduced the notion that it is impure public goods that are provided by SNGs. Second, he demonstrated that mobility of taxpayers – voters – between jurisdictions represents the mechanism by which individuals reveal their preference for ‘impure’ public goods.

2.3 Fiscal Equivalence and the Decentralization Theorem

Olson (1969) introduced the seminal notion of ‘fiscal equivalence’ to economics within a general framework for investigations of fiscal decentralization. ‘Fiscal equivalence’ is the notion which posits that, for every collective good, there is a unique ‘boundary’ for which a separate government is needed, so that ‘there can be a match between those who receive the benefits of a collective good and those who pay for it’ (Olson, 1969, p. 483). In broad terms, it positively associates the efficiency goal of public economics with aligning the costs and benefits of impure public good provision with multi-tiered federal systems, each with overlapping physical boundaries, but each with unique boundaries relating to the provision of specific public goods.

The next seminal contribution was made by Oates (1972) in his monograph *Fiscal Federalism*. Oates implicitly blended Olson’s notion of fiscal equivalence with aspects of Tiebout’s notion of impure public goods in his theory of fiscal decentralization, although without focusing directly on household mobility or sorting. He argued that there should be a variation of the provision of impure public goods and services from governments since inhabitants have different tastes for public services. Oates formalizes his treatment of the issue by defining public goods in a manner that comprises both pure and impure public good attributes. All that matters is that benefits from non-excludable and non-rival public goods are linked to a population in a geographic subset (impure public goods) or the union of population from all such subsets (pure public goods). From this, he developed the decentralization theorem.

For a public good – the consumption of which is defined over geographical subsets of the total population, and for which the costs of providing each level of output of the good in each jurisdiction are the same for the central or the respective local government – it will always be more efficient (or at least as efficient) for local governments to provide the Pareto-efficient levels of output for their respective jurisdictions than for the central government to provide any specified and uniform level of output across all jurisdictions. (Oates, 1972, p. 35)

Oates’s decentralization theorem has a clear rationale. Welfare is maximized when specific public goods are provided by local governments whose jurisdiction corresponds to the subset of the national population for which the demand for specific public goods and services is homogeneous. If the national government is assumed to provide the same bundle of public goods and services across all subnational jurisdictions, then it will not be possible for the national government
to provide the efficient level unless preferences are homogeneous for all members of the population.

Oates’s assumptions on uniform provision of public goods by national governments do not reflect reality in any strict sense. Indeed, this assumption has been criticized by Brennan and Buchanan (1980). However, it may be a reasonable approximation for two main reasons. First, the national government does not have extensive information on a diversity of local preferences and tastes. There is an asymmetry of information relating to local tastes between the national government and SNGs. Second, while national governments can provide different levels or different types of public services in different regions, there are political constraints for the extent to which this can happen. In this regard, the national government is, to some extent, constrained to provide uniform goods and services across regions to avoid possible adverse political consequences which may arise from differentiating the level of goods and services provided across subnational jurisdictions. Related to this are potential constitutional constraints. For example, in Australia, the Federal Government is required to treat states the same – different rates of the Goods and Services Tax could not apply in New South Wales and Western Australia, even if levels of public services demanded in these two states are different.

Oates’s decentralization theorem is also predicated on the two other restrictive assumptions concerning economies of scale and externalities. First, it is assumed that provision costs are the same when public goods are provided by either the national government or all SNGs together. If this is not the case, it is possible that centralized provision may be less expensive when there are economies of scale. However, even this has been criticized by Wagner (2007, pp. 164–167), who argues that economies of scale are largely irrelevant to the size of SNG units. When particular government services benefit from economies of large-scale production, large SNGs enjoying economies of large-scale production can produce these services for sale to other small-scale SNGs. That is, SNGs with diseconomies of scale can purchase public services from larger SNGs which can exploit their economies of scale. As such, some government units articulate demand for services whereas the others produce services for different governments on a contractual basis. Second, there are no externalities in public goods provision – each level of government provides an efficient amount of public goods to a respective constituency. As a result, there is an existence of a so-called perfect mapping – local governments provide goods and services which bring benefits to local people in the same region only. This too is a very ‘strong’ assumption because it is, in practice, almost impossible to limit the benefits of locally public goods provided in a specified region. Externalities always go together with any provision of public goods.

The important contribution of first generation theory is that it reveals that efficient levels of publicly provided outputs are more typically achieved through multi-tiered systems of government than through a unitary system of government. Welfare benefits from decentralization are likely to be greatest when there is a diversity of preferences for impure local public goods. However, the limits of these findings – due to the restrictive assumptions used to derive the decentralization
theorem – cannot be overlooked. At some point, the level of fiscal decentralization will become inefficient as economies of scale for public good provisions are lost and negative fiscal externalities emerge.

2.4 Public Choice and Leviathan

The final stream of the first generation theory derives from the public choice literature, which reaches back into intellectual history for its motivation. In 1660, Thomas Hobbes described the notion of the Leviathan, or to be called a Commonwealth or State, as ‘an artificial man, though of greater stature and strength than the natural, for whose protection and defence it was intended; and in which the sovereignty is an artificial soul, as giving life and motion to the whole body...’ (Hobbes, 1660, p. 1).

Brennan and Buchanan (1980) revived the notion of the State as Leviathan. They developed the hypothesis that the main interest of the government is to tax heavily so that they have financial resources to spend. In this representation, the government is a monolithic Leviathan, which always seeks to maximize its taxation revenue. In response to this issue, the only way to limit the extent of government oversupply, and as a consequence over-taxation, is to constrain governments through effective constitutions that decentralize political and fiscal authority. When political action has the motivational characteristics of Leviathan, political and fiscal decentralization divides, competition between public bodies reduces the force to grow the public sector. In addition, the ‘protective state’, which establishes the government as enforcer of individual rights and contracts, carries the functions which can be effectively allocated, by competition processes, to the relevant level of government. It is hypothesized that if these protective functions are all assigned to the national government, this government unit has a real incentive to maximize a net surplus because there are no effective controls on its taxing powers. The presence of fiscal decentralization constrains government, encouraging it to devolve to increase the efficiency in providing goods and services to its respective jurisdictional inhabitants; otherwise local citizens will vote on their feet. This contributes an effective constraint on excessive taxing from all government units. As a result, the size of the SNGs, and then the size of the aggregate government sector, decreases.

In general, the Leviathan hypothesis and public choices approach to fiscal federalism provided a new perspective on government which highlighted the main advantage of fiscal decentralization as a force for maintaining smaller government.

Total government intrusion into the economy should be smaller, ceteris paribus, the greater the extent to which taxes and expenditures are decentralised, the more homogeneous are the separate units, the smaller the jurisdictions, and the lower the net regional rents. (Brennan and Buchanan, 1980, p. 185)

2.5 A Brief Characterization of the First Generation Theory

In broad terms, it may be concluded that there are two general streams of first generation theory, with Tiebout having a major influence on both. There are
(i) studies that draw on Tiebout’s impure local public goods concept and integrate it with the Musgravian framework and (ii) studies that draw on Tiebout’s notion of inter-jurisdictional mobility and link it with forces that limit the size of the public sector. The work of Oates, for example, would fall under the first category, referred to here as the ‘core’ first generation theory of fiscal decentralization, while the public choice approach of Brennan and Buchanan would fall under the second, non-core, category. Importantly, though, the non-core public choice approach is a ‘complement’ to ‘core’ first generation theory, as it relates to the particular question of fiscal decentralization.

3. Core First Generation Approaches to Fiscal Decentralization

3.1 The Assignment of Service Delivery Responsibilities

The economic boundary of individuals’ net benefits (benefits less costs) from the provision of public goods may extend beyond the political boundary of the host jurisdiction, in which case positive fiscal externalities are provided to residents of neighbouring jurisdictions. The economic boundary of individuals’ net benefits from public goods may also fall within the formal political boundary of the government which provides them, in which case some residents within one jurisdiction receive a benefit which exceeds the costs they pay, while other residents within that jurisdiction receive a benefit that is less than the costs they pay. Finally, the economic boundary of a public good matches the political boundary of the government that provides it. Prima facie, only the final relationship reveals the presence of economic efficiency in providing public goods; the first two cases are inefficient.

‘Core’ first generation theory of fiscal decentralization reveals the above relationship well. However, it does not represent a fully deterministic system for the efficient assignment of responsibilities across various levels of government for the provision of particular services. Only general ‘principles’ are evident from the efficiency analysis under the first generation theory, with the notion of ‘fiscal equivalence’ being an especially important concept. In general, five main principles of fiscal decentralization may be deduced from ‘core’ first generation theory.

First, pure nationally bounded public goods are more efficiently provided by the national government (Bird, 2000). For example, foreign policy, defence, immigration and international trade can be best formulated and implemented by the national government. Services should be provided by the national government when demand is at a constant level across the various subnational localities. In addition, centralized provision of public services also experiences a great benefit from economies of scale.

Second, based on the principle of fiscal equivalence and Oates’s decentralization theorem, the geographic boundaries of the diverse range of local public goods should align (as far as possible) with political boundaries of the second-tier governments (states) and the third-tier governments (local governments). SNGs are able to provide a range of services to local communities such as law, order
and public safety, education, health policy, as well as very local issues such as the street lighting system, local sewerage, garbage collection, and local paper deliveries. Importantly, variations in the level of provision of public services across subnational regions provide a basis for partially redressing (i) spill-over effects and (ii) congested effects.

**Third**, public provision of both pure ‘private’ goods and impure SNGs’ public goods and services should be based on the size of jurisdiction, and in accordance with local tastes and preferences (Shah, 2004). If the size of jurisdiction is considered, the principle of benefit matching is achieved because local citizens who receive benefits also bear costs. A system of fees and user charges may also be useful and effective for the purpose of cost recovery (McLure and Martinez-Vazquez, 2004). SNGs operate closely to local inhabitants so that they are the sole agents, who are in the best position to understand preferences, tastes and the amount demanded. Once the ‘benefit areas’ can be established, local provision on the basis of cost recovery tailors local service provision to the demands of local people. This enhances economic efficiency. Related to this, local provision of public services may also result in experimentation, and then innovation, to promote efficiency in public policy for the entire economy (Oates, 1999). This suggests that devolution may also have ‘dynamic’ efficiencies.

**Fourth**, the assignment of responsibility for the various types of service delivery must be transparent and clearly understood and agreed by all parties. Failure to do so results in the overlapping of publicly provided services. Clear assignment of responsibility for service provision limits co-sharing responsibilities, where more than one level of government gets involved in the same areas of spending (Martinez-Vazquez, 2001). Co-sharing, while sometimes unavoidable, can lead to ambiguity which creates unnecessary coincidence in providing public services to local communities, and, in turn, negatively affects the efficiency of the spending programmes. It may even be used to intentionally confuse the responsibilities of each level of government. This may result in fiscal illusion – local citizens may misjudge the ‘true’ benefits and costs of their government (Dollery and Worhington, 1999). Two effects associated with the notion of a citizen’s fiscal illusion are widely recognized (Turnbull, 1998). First, public spending is maintained at a greater level under fiscal illusion compared with perfect information – the overspending effect. Related to this is the ‘flypaper effect’, which is the prediction that intergovernmental grants from high level government usually stimulate more local spending than locally generated revenue. Clear assignment of responsibility alone does not necessarily overcome the problem of fiscal illusion.

**Fifth**, the economies of scale of local production of goods and services, including the related issues of local financial, managerial and administrative capabilities, should also be considered. This is particularly important for developing economies. As such, asymmetric spending assignment may be appropriate. That is, service delivery responsibility may be different across the same level jurisdictions when there are different economies of scale and administrative capacities across governments within the same level of SNGs. However, in reality, countries

© 2009 Blackwell Publishing Ltd
which adopt this type of spending assignment usually do so for political and/or geographical reasons (Joumard and Kongsrud, 2003).

In short, the first generation theory of fiscal decentralization suggests that, in general, service provision responsibilities are best assigned to the lowest level of government that can meet the service responsibility efficiently (Bird, 2004). That is, the so-called ‘law of subsidiarity’ is a simple summary of first generation theory. The law of subsidiarity states that economic performances of the government become more productive, efficient and effective if public services are provided by the lowest level of government possible (Martinez-Vazquez, 2001, p. 12). Economic efficiency and accountability are essential features for SNGs’ fiscal autonomy in terms of service delivery. As such, efficiency and accountability are strongly linked to the autonomy of SNGs in fiscal matters, subject to the law of subsidiarity – first generation theory supports SNGs having the autonomy to decide the spending programmes within their own jurisdiction. This also minimizes the national interference into local spending programmes through the requirements of SNGs’ accountability to those who pay taxes and charges. The efficiency of economic activities of SNGs is also enhanced because SNGs are in a better position to know where the money should be heading.

3.2 The Assignment of Revenue-raising Powers

The potential benefits of ‘efficient’ tax assignment have been an important subject for consideration in first generation theory of fiscal decentralization. In terms of the degree of decentralization, the following assignments of revenue-raising powers represent a progressively decreasing level of SNGs’ fiscal autonomy (OECD, 1999): (i) SNG sets tax rate and tax base; (ii) SNG sets tax rate only; (iii) SNG sets tax base only; (iv) tax sharing arrangements with two typical different levels – SNG determines the revenue split or the national government determines the split – and (v) national government sets rate and base of SNG tax.

In an influential paper, Musgrave (1983) identified six fundamental principles for tax assignments: (i) taxes suitable for economic stabilization should be national; (ii) taxes with unevenly distributed bases across jurisdictions should be central; (iii) personal taxes with a mobility of tax bases should be assigned to the national government; (iv) taxes with cyclically stable revenue should be assigned to SNGs; (v) SNGs should be assigned taxes with bases that are not easily mobile among jurisdictions; and (vi) benefit taxes, users’ fees and charges are relevant to any level of government. These six principles are discussed in turn below.

The first three principles concern taxes that should be assigned to the national government. Based on Musgrave’s contention that there are three branches of public finance (stabilization, distribution and allocation efficiency), fiscal decentralization is primarily justified in terms of allocation efficiency. The first principle reaffirms this by assigning the economic stabilization branch to the national government because it has responsibility for macroeconomics policy. Moreover, subnational stabilization policies are likely to compromise stability-oriented policies by ‘leakage’ outside individual SNGs, which would effectively diminish the national
government’s capacity to control aggregate demand. As such, personal income tax and corporate income tax are suitable for national governments – especially when income redistribution is a national goal. However, it is also argued that a proportional income tax without any significant distribution goal can be a desirable source of revenue for SNGs because SNGs may decide to piggyback on the national income tax. In addition, together with the distributional issue, implementation issue is also an argument for the inappropriateness of the corporate income tax at the subnational level. The second principle is related to unevenly distributed taxes such as the natural resources taxes. Natural resources are unevenly distributed in some particular jurisdictions. Political and social tensions can be avoided if revenue from these sources is centrally generated and redistributed to all regions. If not, jurisdictions with national resources may decide to provide local services at a lower revenue-raising burden. The third principle is related to revenue sources with mobile bases. The concern is that tax rate differentials between SNGs will cause inefficient movement of such bases across borders.

Principles four and five concern tax assignment fundamentals necessary to support SNGs. The fourth principle is designed to secure stable long-term sources of revenue for SNGs. Because of their relatively small fiscal size, SNGs’ need for relatively stable revenue flows is greater than that from the national government. Stable revenue sources also assist with continuity of public service provision. The fifth principle is essentially the converse of the third principle – taxes with volatile mobile bases should be national, and tax bases that are immobile and stable should be managed by SNGs. Property and payroll taxes are two good examples. There is a balance, however, that this principle must strike with the second principle – for example, natural resources tax (tax with unevenly distributed bases) should be with the national government even when they generate stable revenue flow. The property taxes are obviously immobile but they may also be unevenly distributed across SNGs.

The sixth principle concerns common types of government fees or charges which are efficient irrespective of whether imposed by the national government or SNGs. Each level of government can employ charges that link directly to the services provided. For example, driving licence fees and petrol tax can be used to finance a highway, but not for education. When services benefit all people, this type of charge is best administered by the national government. However, when the benefits are only directed to a particular group in a particular region, SNGs would be better placed to administer the charges.

All these principles are designed to ensure that taxation and revenue-raising powers across all levels of government do not distort economic choices of non-government entities and individuals. However, they are also derived from the static welfare analysis – tax competition between jurisdictions is generally not supported because, of necessity, it distorts private decisions. There is no clear and explicit consideration given to the issue of whether the dynamic benefits for fiscal competition between SNGs generally, and tax competition specifically, outweigh the ‘static’ welfare losses from the distortions to individual decisions resulting from tax competition.

© 2009 Blackwell Publishing Ltd
3.3 Intergovernmental Fiscal Transfers

The assignment of responsibility for service deliveries and taxing powers across the various levels of government is the first-order (or constitutional level) issue in first generation theory. Fiscal transfers between levels of government are essentially a second-order (or non-constitutional level) issue in the sense that they are conditioned by the degree of vertical fiscal imbalance under the prevailing assignment of revenue and expenditure responsibilities (or implied by the constitution). Nevertheless, the literature on fiscal federalism that investigates fiscal transfers, usually from the national government to lower tiers, is extensive and detailed because the potential for formalized analysis of fiscal transfer issues is greater than in the case of the more fundamental issues concerning the assignment of fiscal powers.

The essential point concerning fiscal transfer in first generation theory is that transfers are either ‘conditional’ (tied) or ‘unconditional’ (untied) and that each has different consequences for the fiscal autonomy of SNGs (Oates, 1972). While unconditional grants provide a higher degree of fiscal autonomy of SNGs, conditional grants actually limit this autonomy. Fiscal autonomy of SNGs in terms of fiscal transfers is generally enhanced when the majority of grants from national government to SNGs are in the form of unconditional grants. That is, unconditional grants are generally supported by first generation theorists ahead of ‘conditional’ grants (Bird, 2004) unless tied grants are required to correct for market failure or inter-jurisdictions spill-over effects.

Boadway (2003) argued that while fiscal equity is matched for unitary countries when their social preferences are the same across regions, the notion of fiscal equity is more or less ambiguous in federal countries because national governments and SNGs use a mix of fiscal policies to achieve a mix of redistribution goals. Every citizen is exposed to at least two distinct sets of fiscal distribution – one from the national government and another from SNGs. When there are differences in taxing capability between jurisdictions, and, as a result, differences in provision of services to the given standard among different regions, horizontal fiscal imbalance arises. Imbalance in the fiscal capacity of SNGs at the same level is associated horizontal fiscal inequity. This phenomenon can be caused by differences in local wealth, natural resources, per capita tax capacity, and differences in spending obligations and in the cost of providing public services (Martinez-Vazquez, 2001). The main tool for achieving equity through intergovernmental fiscal transfers is the mechanism ‘fiscal equalization’. Fiscal equalization is a process which reduces fiscal disparities among subnational regions through financial transfers (Collins, 2001). The four main reasons for fiscal equalization are identified in Schroeder and Smoke (2003): (i) to fill the fiscal vertical gap between national government and SNGs in aggregate; (ii) to equalize horizontal fiscal capacity for governments in the same level; (iii) to internalize fiscal externalities within SNGs and (iv) to correct the administrative weaknesses of SNGs.
3.3.1 Fiscal Equalization

The first generation theory of fiscal equalization is the most developed strand of the economics of fiscal decentralization. Perhaps the most formal developments in this field concern the question of fiscal equalization arrangements.

Economists first came to this issue of fiscal equalization because of a concern over resource misallocation resulting from one jurisdiction, say a ‘rich’ state, subsidizing another jurisdiction, say a ‘poor’ state, through equalization arrangements. A seminal study in this regard is ‘A note on grants in federal countries’ by Scott (1950), the specific concern being that, as labour is mobile, fiscal equalization will have the effect of inducing residents to low productivity (i.e. poor) jurisdictions and away from high productivity (i.e. rich) jurisdictions, causing a decline in the value of national output. However, fiscal transfers from one government to another have some important differences from the case where governments subsidize private enterprises. First, fiscal equalization is, in effect, a transfer between states where all states pursue equity goals within their jurisdiction by providing some individuals with a positive fiscal residuum or net fiscal benefits (value of public benefits consumed less taxes paid to some individuals) and a negative fiscal residuum or negative net fiscal benefits to other individuals. Second, in addition to equity issues, the ‘local’ character of subnational public goods has potentially adverse congestion effects in very populous and highly productive (rich) jurisdictions which can reduce the value of public goods in those jurisdictions. Moreover, services provided by SNGs may have benefits (or costs) that are not confined to the geographic area of the jurisdiction that provided the service. Once these ‘fiscal externalities’ factors were recognized, it was no longer obvious that the ‘subsidies’ from one (or many) government(s) to other governments necessarily lead to a misallocation of resources. Under certain circumstances, such transfers may even be efficiency enhancing.

Perhaps the first true classic study in this field after Scott was ‘Federalism and fiscal equity’ by Buchanan (1950). Although equity is the focus of the proposition that ‘like individuals’ in different jurisdictions should receive ‘like treatments’ (Buchanan, 1950, p. 588), it has an important implication for efficiency too. If all jurisdictions of a given level provide greater positive net fiscal benefits to the poor, but do so to different degrees, then inefficient migration will result: rich citizens will have an incentive to relocate to the jurisdiction where their negative net fiscal benefit will be the least and poor citizens will have an incentive to relocate to jurisdictions with the highest positive net fiscal subsidy for the poor. If fiscal equalization transfers result in the like treatment of like individuals across different subnational jurisdictions in a federation, then ‘like’ rich people will face the same negative net fiscal subsidy in all jurisdictions, and ‘like’ poor people will also face the same positive net fiscal subsidy in all jurisdictions. As such, there is no fiscally induced incentive resulting for intergovernmental transfers to distort migration. Of course, this is predicated on the view that all SNGs use fiscal powers to achieve the same vertical fiscal equity goals (i.e. redistribution of income).
Buchanan’s equity approach sets aside the influence of congestion in local public services and the effects of spill-overs. These issues were investigated more by scholars, such as Buchanan and Wagner in ‘An efficiency basis for federal fiscal equalisation’. Buchanan and Wagner pointed out that, for any constant level of public facility, an addition of one more consumer will reduce the quantity of consumption units available to other people (Buchanan and Wagner, 1980, p. 244). This is because local public goods may be subject to congestion. Furthermore, given the public good character of service provision, people who move to the high productive (i.e. richer) state ‘exerts an external economy on residents of that state and an external diseconomy on residents of the state that he leaves’ (Buchanan and Wagner, 1980, p. 245). However, when SNGs provide impure public goods, a point is reached where the highly productive state faces an external diseconomy – the benefit to all residents from tax-side externalities (i.e. the reduced average cost of public good provision) is swamped by congestion costs. In such circumstances, fiscal equalization may enhance efficiency by preventing inefficient migration from poor to rich states. The important contribution of Buchanan and Wagner is that they identify the circumstances when, from a pure resource allocation perspective, fiscal equalization is economically efficient.

In addition to the fundamental issues of congestions and other fiscal externalities such as spill-overs (i.e. economic consequences of the activities of an SNG that spill outside the border of that SNG), a number of other important ‘complications’ have emerged in the theory of fiscal equalization. These include the efficiency implications including the correct treatment in fiscal equalization processes of economic ‘rents’ (Boadway and Flatters, 1982, p. 621) and differences in the ‘unit cost’ of services in different SNGs. When economic rents from mineral activities differ between states and they are taxed by an SNG that is subject to fiscal equalization, should those rents collected by the SNG be equalized (effectively allocating rents on a per capita or some other basis across all equalized subnational jurisdictions) or should the SNGs in which the rents are generated keep the taxes from that rent? Likewise, in SNGs where the unit costs of inputs required to produce local public services are above those for other SNGs, should the high-cost SNGs be subsidized for their higher unit costs through the fiscal equalization process?

In regard to rent analysis, Boadway and Flatters (1982) demonstrated that fiscal equalization represents an efficiency enhancing mechanism. Individuals will choose to reside in the jurisdiction that provides them with a highest net benefit, but when there are location specific rents, migration equilibria will not equate with the point where marginal products are equal across jurisdictions (i.e. the efficient point) but where average products are equal. As in the case of congestion, fiscal equalization provides one mechanism by which migration distortions due to location specific rents can be eliminated.

Petchey (1995) has also investigated the interaction between equalization in the presence of both rents and unit cost differentials. His economic simulations illustrate that, in the presence of diverse economic rents and diverse unit costs across regions, fiscal equalization processes which equalize for unit cost can enhance economic
efficiency. Based on a simplified two states model, he demonstrated that, in the absence of economic rents, efficient fiscal equalization will not transfer resources from the ‘low’ cost state to the ‘high’ cost state. However, when a state has high costs in conjunction with low economic rents (all determined relative to the other states), efficient fiscal equalization transfers will transfer resources from the low cost/high rent state to the high cost/low rent state.

Finally, it should be noted that there is a basic issue concerning what exactly should be equalized – the capacity of governments for providing services (i.e. fiscal capacity) or the actual fiscal performance of a government (i.e. fiscal performances). The literature tends to implicitly focus on fiscal performance (for example Buchanan, 1950; Scott, 1950; Buchanan and Wagner, 1980) but in practice it is generally fiscal capacity that is equalized.6

3.4 Subnational Borrowings

SNGs’ access to the financial market has been a controversial issue for some time – especially in the literature relating to developing countries (Ter-Minassian, 1996). Compared with the national government, some argue that SNGs have poor accounting and disclosure standards and rely on the national government to bail out a debt crisis when there is a conflict of interests between lenders and borrowers. However, there are three main grounds to support SNGs’ borrowings (World Bank, 2004). First, SNGs’ investment expenditure entirely financed from current tax revenue is inappropriate for equity reasons. Investment brings benefits for future generations so that they should be, at least partially, responsible for their spending. Second, the mismatch between revenue and expenditure evident in most federations also requires SNGs’ borrowings in particular years as subnational economies fluctuate. Third, and most importantly from the perspective of first generation theory, borrowings from SNGs may foster political accountability. In the absence of significant intervention from the national government, financial markets will send a clear signal of SNGs’ performances by providing borrowings or blocking them access into the financial markets. As the first approximation in which inter-jurisdictional fiscal externality is ignored, first generation theory suggests that SNGs should have some responsibility for the management and service of their own debt, but in conjunction with transparent accounting arrangements for public sector finances.

In practice, one of the main questions of interest to scholars is the extent of the budget constraint to be imposed on SNGs by a national government (Ter-Minassian, 1996, p. 156): a soft budget constraint and a hard budget constraint. A hard budget constraint means that the national government will not bail out SNGs for excessive debts under any circumstances. This approach requires the national government to provide a clear message to the market – no bailout policy for SNGs’ debts. In this case, SNGs have no ‘incentives’ to over-borrow and fail to meet financial responsibility. On the other side, soft budget constraint means, to some extent, subnational debts are implicitly guaranteed by the national government. This approach has usually been adopted in countries with high levels of fiscal transfers.
In this case, local regions are very much dependent on outside-sourced revenue, not their own-sourced revenue.7

4. An Emerging Second Generation Approaches to Fiscal Decentralization

Oates’s decentralization theorem has laid a strong foundation for the study of fiscal federalism and decentralization, all of which developed from the general literature on public economics. The result was what is now called the first generation theory of fiscal decentralization. However, towards the end of the last decade of the twentieth century, a ‘second’ generation theory of fiscal decentralization has begun to emerge, which draws on ideas from outside the public finance literature. As Oates (2005, p. 356) has highlighted, this second generation began investigating fiscal decentralization by drawing on notions from the theory of the firm, the economics of information, the principal–agent problem, and the theory of the contract.

Two main considerations underlie the development of the second generation theory (Oates, 2005, p. 356). The first concerns the political processes and the behaviour of political agents in which participants may have their own objective functions. Government officials may not need to seek the common good as assumed in the first generation theory; rather, they may not act to maximize the welfare of their constituencies. This consideration has obvious links to public choice theory – which was the main ‘non-core’ stream of the first generation theory of fiscal decentralization. The second concerns the issue of asymmetric information and political agents. Some particular participants have more knowledge of local preferences and tastes and cost structure compared to the others. To examine these influences, fiscal federalism is examined from the perspective of a framework on industrial organization and microeconomic theory. Much of this work by the second generation theorists concerns the issue of balance between the degree of fiscal centralization and fiscal decentralization. While the general support for fiscal decentralization in the first generation theory is acknowledged, the dangers of going too far in the fiscally decentralized system are a feature of the second generation theory.

The emerging second generation theory has been characterized in terms of two motivating issues: incentives and knowledge (Vo, 2008). Both motivations have contributed to an increased economic efficiency: incentives are required for SNGs to do a better job to avoid outward migration of people and firms; and knowledge of local preferences and tastes is crucial to achieve economic efficiency when local public goods and services are provided by SNGs. The contributions of the second generation theory are mainly drawn from the economics of transaction cost, incomplete contracts and principal–agent perspectives (Vo, 2008). Leading studies, that have been classed as parts in the emerging second generation theory, are associated with Weingast (1995), Seabright (1996), Lockwood (2002) and Besley and Coate (2003).

Weingast (1995) introduced the notion of market preserving federalism to investigate how competing jurisdictions create incentives for credible commitment.
and lower transaction costs. Assuming a hierarchy of governments (at least two levels of government rule the same land and people) with autonomy of each level of government institutionalized, Weingast modelled the market preserving federalism in which (i) SNGs have regulatory responsibility over the economy, (ii) a common market is ensured so that SNGs are unable to prevent trade in goods and services from other jurisdictions and (iii) SNGs face a hard budget constraint. The main findings, for the UK and the USA are (i) federalism provides the political basis for the common market; (ii) the prohibitions against the national government’s exercise of economic regulation greatly reduced the government’s political responsiveness to interest groups; and (iii) the prohibitions on internal trade barriers allowed entrepreneurs, new enterprises, and new economic activities to emerge in new areas that could outcompete interests in older areas (Weingast, 1995, p. 25).

In contrast, Seabright (1996) introduced the notion of ‘incomplete contract’ to the analysis of fiscal federalism. The author presents elections as incomplete contracts in which some information, in the ‘contract’, is unverifiable. Political accountability can be an organizational motivation for decentralization. In contrast, centralization could be more preferred when the mechanisms associated with incomplete contract provide greater scope for policy cooperation between different levels of government to internalize inter-jurisdictional fiscal externalities. As a consequence, the decision on preferred mechanism depends on the relative magnitude between benefits from internalization of inter-jurisdictional fiscal externality and costs arising due to a reduced accountability under fiscal centralization.

The starting point for Lockwood (2002) and Besley and Coate (2003) is Oates’s decentralization theorem, but these authors correctly pointed out that goods and services provided by the national government are not necessarily homogeneous, as Oates had originally assumed. Once it is recognized that national provision of public goods and services is possible on a differential basis between regions, a different fiscal framework is needed from that developed by Oates. In their frameworks, output provided by the national government consists of locally designed outputs which are determined by the central legislation. This is feasible because the national government body always consists of locally elected representatives from local regions. Nevertheless, the broad thrust of Oates’s findings on the efficiency of decentralized fiscal arrangement is reproduced in models in which inter-jurisdictional externalities are small and regions are heterogeneous.

Another contribution to second generation theory of fiscal decentralization is evident from recent studies of Australian fiscal equalization by Petchey and Levchenkova (2002, 2004), which pointed out that equalization grants are not determined exogenously. This is because equalization is undertaken with respect to a fiscal standard that is determined in relation to actual fiscal behaviour. In this circumstance, there is an incentive for the states to behave strategically in their fiscal decisions. This provides the basis for an ‘equalization game’ which takes into account states’ strategic behaviour and provides a basis for developing proposals which enhance efficiency.
The generally accepted view now is that the presence of the second generation theory of fiscal decentralization classifies, rather than contradicts, the validity of the first generation, including the decentralization theorem:

Although the models under the second generation theory umbrella differ in fundamental ways from the first generation theory, many of them produce a trade-off between centralisation and decentralisation that is in a somewhat similar spirit to their earlier counterparts. (Oates, 2005, p. 357)

Finally, a second generation of public choice theory is also emerging. This is mainly associated with the work of Wagner (2007), which shifts attention away from ‘tiers’ of government, and assignment of powers and responsibilities across tiers, to focus on the competing ‘polycentric’ character of government. This is associated with a focus on the process by which diverse centres of government respond to demands for publicly provided services in a manner that is integrated within the market economy.

Wagner’s basic view on the evolutionary character of polycentric government is more important than the hierarchical aspect of government in which the questions of which functions to be provided at the national and subnational levels of government are addressed (Wagner, 2007, pp. 164–165), as it is in the conventional approach to fiscal decentralization. The process by which governments respond to emerging demand from the community is directly linked to the capacity to raise their own revenue in an innovative way. Moreover, the focus is on government ‘enterprise’, in which some government units fulfil the role of public service producers and some take on the role of the articulators of the public services (Wagner, 2007, p. 166).

5. Concluding Remarks

The paper provides a general and brief overview of the main insights from the literature on fiscal decentralization with the focus exclusively on theoretical developments of fiscal federalism and decentralization. It reveals the findings from the first generation theory, as represented by seminal studies from Musgrave, Tiebout, Olson and Oates, ‘fiscal equivalence’, population sorting and the fiscal decentralization theorem. It also points to the public choice stream of first generation theory and an emerging second theory of fiscal decentralization, which mainly concerns the efficiency trade-off between fiscal centralization and decentralization.

The key point to note is that the notion of decentralization plays a very important role in the theory of fiscal federalism; irrespective of whether core or non-core first or second generation theory is being considered, decentralization is the primary issue of concern. For the purpose of this paper, one important implication of this overview is that any empirical testing of this theory requires a ‘sound’ measure of fiscal decentralization. That is, there must be a measure of fiscal decentralization that reflects important features of fiscal theory.
Acknowledgements

I am greatly indebted to Michael McLure and Kenneth Clements for the support and constructive comments they have provided during the writing of this paper. I acknowledge financial assistance from an International Postgraduate Research Scholarship from the Australian Government, Postgraduate Award from the University of Western Australia and the School of Economics and Commerce, UWA. Dr Duc Vo, formerly at the University of Western Australia, is currently working at the Economic Regulation Authority in Western Australia, Australia.

Notes

1. The origin of the third type of argument derived from public choice theory – especially the notion of fiscal illusion and the flypaper effect (Turnbull, 1998).

2. The assignment must not generate inefficient signals in which one level of government has an incentive to oversupply or undersupply services. Healthcare provision in Germany is an example. In Germany, the health fund from the national government is responsible for financing the costs of hospitals. The Landers (the first level of SNGs) finance hospital investment and decide hospital capacity. The municipalities (the second level of SNGs) are assigned to provide local healthcare services. In this context, the Landers have an incentive to expand hospital capacity to create local employment whereas the operating costs have to be met by the national government (Wurzel, 1999).

3. A referee argued that the OECD has missed the first most important form of tax autonomy: whether subnational governments have discretion of not to introduce their own tax.

4. Of course, there are other ways for SNGs to manage the revenue volatility issue. Governments could ‘sell’ volatile revenue flow to a buyer who was prepared to exchange a stable revenue flow. Alternatively, cyclical accounts could be established which are drawn down by SNGs in recessions, and added to during booms.

5. Parry (2003) shows that competition among various subnational governments may lead to a suboptimal level of capita taxation when governments ignore the ‘external benefits’ of capita flight to other regions.

6. Mathews (1980), Musgrave (1980) and Zabalza (2003) have distinguished five main types of models of fiscal equalization: (i) the fiscal capacity equalization; (ii) the fiscal potential equalization; (iii) the pure incentive model of fiscal decentralization; (iv) the actual performance equalization and (v) the Mathews model. The emerging second generation theory of fiscal decentralization considers some aspects of capacity equalization. First, Mathews raised two main aspects of fiscal capacity equalization: equalization of revenue-raising capacity and equalization of expenditure needs, similar to that which is applied in Australia (see Section 2.3.3). Second, the fiscal potential model equalizes the fiscal potential levels for SNGs to achieve their own goals). Third, the pure incentive model of fiscal equalization is done to give an incentive to subnational regions by providing matching grants to them to raise their tax and service levels. This model of fiscal equalization may have redistributional effects with respect to fiscal capacity and/or fiscal need of various regions. Fourth, the equalization is demonstrated by the national government by equalizing actual per capita dollar outlays on subnational services in all subnational regions and taking into account the difference of needs of various regions. Fifth, the Mathews model has
developed on the foundation of the fiscal capacity model of fiscal equalization and explored various specific situations, such as taking into account (i) a revenue effort adjustment factor or (ii) a revenue capacity equalization factor.

7. In order to limit a possible weakness of the guarantee for subnational debt repayment, many restrictions have been imposed such as no borrowings, borrowings for limited purposes, and ceiling amount control (Rodden, 2000). While these soft constraints are moderately flexible, the approach experiences the problem of the commons. The commitment to bailouts for subnational debts by national governments may bring fiscal benefits for SNGs as an entity; however, it need not benefit the residents of subnational regions when this is achieved by reducing the financial accountability of SNGs and damaging the national economy. It is very costly, however, for the economy as a whole.

8. It should be noted that public choice economists such as Brennan and Buchanan had previously objected to this.

References


Boadway, R. (2003) National taxation, fiscal federalism and global taxation. The WIDER Project on Innovation Sources of Development Finance, Queen’s University, Canada.


