ISSN 0732 507X

PN-ABF-568
67108

OCCASIONAL PAPER NO. 138

RETURNS TO SCALE IN PUBLIC PRODUCTION

William D. Duncombe and John Yinger

METROPOLITAN STUDIES PROGRAM
THE MAXWELL SCHOOL OF CITIZENSHIP AND PUBLIC AFFAIRS
SYRACUSE UNIVERSITY
SYRACUSE, NEW YORK 13244-1090

June 1990

$5.00

SYRACUSE UNIVERSITY - SPRING 1990

Melvin A. Eggers — Chancellor

THE MAXWELL SCHOOL

John Palmer — Dean

THE METROPOLITAN STUDIES PROGRAM

David Greytak	Economics	Director
Michael Wasylenko	Economics	Associate Director
Larry Schroeder	Public Administration/Economics	Director DFM Project

SENIOR RESEARCH ASSOCIATES

Nancy Benjamin	Economics
Guthrie Birkhead	Public Administration
Stuart Bretschneider	Public Administration
Beverly Bunch	Public Administration
Jesse Burkhead	Economics/Public Admin.
James Follain	Economics
Ali Galaydh	Public Administration
Vernon Greene	Public Administration
William G. Johnson	Economics
Bernard Jump, Jr.	Public Administration
Mary Lovely	Economics
Jerry Miner	Economics
Jan Ondrich	Economics
John Palmer	Public Administration/Econ.
Bruce Riddle	Academic Computing Specialist
Seymour Sacks	Economics
John Schnell	Economics
Thomas Selden	Economics
Jeffrey Straussman	Public Administration
John Yinger	Public Administration/Econ.

GRADUATE RESEARCH ASSOCIATES AND ASSISTANTS

Robert Carroll	Economics
Patricia Coleman	Social Science
Mary Daly	Economics
James Dearborn	Public Administration
Dan Gerlach	Public Administration
Lauria Grant	Economics
Nigel Grant	Economics
Robert Guell	Economics
Maurice Harris	Public Administration
Amy Hamburg	Public Administration
Jocelyn Johnston	Public Administration
Karl Knapp	Public Administration
Jing Li	Economics
Susan MacPhee	Public Administration
Ashley Probart	Public Administration
Jinghu Oh	Economics
Maria Rendon	Public Administration
Taryn Rounds	Public Administration
John Ruggiero	Economics
Gyan Sinha	Economics
Schuyler Tilly	Economics
Kurt Thurmaier	Public Administration
Laura Wheeler	Economics
Sung-Kull Yoo	Economics

STAFF

Margaret Austin	Administrative Assistant
Martha Bonney	Word Processing Specialist
Esther Gray	Administrative Secretary
Laura Sedelmeyer	Secretary
Henry Emmans	Librarian

FOREWORD

A substantial body of discussion has developed about issues of scale economies in local government production. To a large extent, this literature has been concerned with empirical estimation, but it has yet to resolve the issue of whether there are economies or diseconomies of scale in the production of local public goods. In this paper, the authors review the existing literature and find it wanting in several important respects. Most particularly, they point out that all previous studies are based on the assumption of a constant quality level per capita with scale economies attributed to variations in the size of population served. The authors take a novel approach by differentiating between returns to quality scale and to population scale. The value of this approach is that it allows for the separate estimation of the response of government costs to the level of government activity and to population. The results of their empirical analysis of paid fire departments in New York State collaborate some results from earlier studies in that they find constant returns to population scale. In addition, they find significantly increasing returns to quality scale. As Duncombe and Yinger point out, their findings suggest that consolidation of fire departments may not result in cost savings and that increasing the quality of fire protection production may significantly lower the cost per unit of quality.

This paper grows out of work initiated while Bill Duncombe was in residence at the Metropolitan Studies Program as a graduate student in Public Administration. He is currently an Assistant Professor of Political Science at the University of Georgia. John Yinger is Professor of Economics and Public Administration and Research Associate in the Metropolitan Studies Program.

David Greytak
Director
Metropolitan Studies Program
May 1990

TABLE OF CONTENTS

		Page No.
1.	Introduction	1
2.	Defining Returns to Scale in the Public Sector	1
	A Model of Public Sector Production and Costs	3
	Defining Returns to Scale	6
3.	A Review of the Literature on Public Returns to Scale	9
	Technical Returns to Scale	10
	Returns to Population Scale	10
	Returns to Quality Scale	13
4.	Estimating Returns to Scale in Public Production	14
	The Limits of Expenditure Models	14
	A Cost Model for Estimating Public Returns to Scale	17
5.	Estimates of Returns to Scale in Fire Protection	21
	Measures and Data Sources	21
	Empirical Results	23
6.	Conclusions	27
	References	29

LIST OF TABLES

No.	Title	Page No.
1	HYPOTHESIS TESTS RELATED TO RETURNS TO SCALE	24
2	ESTIMATED COEFFICIENTS FOR THE TRANSLOG COST (AND LABOR SHARE) MODEL FIRE PROTECTION IN NEW YORK STATE	26

RETURNS TO SCALE IN PUBLIC PRODUCTION

1. Introduction

Both positive and normative analyses of public spending require an understanding of the technology of public production. Yet despite an extensive literature on the topic, several features of public production remain poorly understood. One such feature is returns to scale. Several different definitions of returns to scale in public production have been proposed and, as we will see, one universal assumption about returns to scale has virtually never been tested.

The purpose of this paper is to present a systematic analysis of returns to scale in the public sector. In section 2 we present a general model of public production and use it to define returns to scale. This framework is then used, in section 3, to review the diverse literature on this topic. In section 4, we develop a specific public-sector cost function that can be used to estimate returns to scale. We apply this cost function to data for fire departments in New York State. Our measures, data sources, and results are presented in section 5. The last section summarizes our key results and discusses their policy implications.

2. Defining Returns to Scale in the Public Sector

In the theory of the firm, returns to scale refer to the change in output that occurs when all inputs are increased proportionately. If output doubles when all inputs are doubled, for example, a technology is said to exhibit constant returns to scale. Increasing (decreasing) returns to scale exist when doubling the inputs more than (less than) doubles output. If factor prices are independent of output, that is, exogenous, and firms maximize profits, then the duality between production and cost functions implies that returns to scale are also revealed in the shape of a

firm's average cost function. Constant average costs, correspond to constant returns to scale, whereas decreasing (increasing) average costs correspond to increasing (decreasing) returns to scale. According to standard terminology, economies (diseconomies) of scale correspond to decreasing (increasing) average costs and are therefore equivalent, when the duality results hold, to increasing (decreasing) returns to scale.[1]

Attempts to employ these concepts in the study of public production have encountered two serious obstacles. First, public outputs typically are difficult to measure, so that public production relationships are difficult to observe. Public expenditures are easily observed, but as we will see, they cannot be incorporated into models of public production without restrictive assumptions about technology. Second, the notion of "scale" in public production has two fundamental dimensions: the quality of the services provided and the number of people served. The former dimension most closely resembles the notion of returns to scale in the theory of the firm, but the latter dimension has received most of the attention. To keep these two dimensions separate, we use separate terms, to be defined more fully below, namely returns to quality scale and returns to population scale.

Virtually all studies of local production or local expenditure deal with the quality dimension by assuming constant returns to scale and exogenous factor prices. As we will see, these assumptions are convenient, in that they lead to single-equation models of public spending, but they also leave untested a key aspect of public production. Moreover, the existing studies of the population dimension yield varied, and sometimes conflicting results, even for the same public service. The purpose of this article is to clear up some of this confusion.

---

[1]See Varian (1984, Ch. 1) for a presentation of the theory of the firm using the dual approach.

A Model of Public Sector Production and Costs

The model of public production presented here borrows from the simple, but powerful conceptual framework developed by Bradford, Malt, and Oates (1969). This framework is built on the important distinction between the direct services provided by a government and the outcome of interest to voters. In the case of fire protection, for example, voters care about the saving of lives and property, not about the number of fire companies available, per se.

Thus, the public production process can be divided into two stages. The first stage of the process is similar to production for a private firm. Local governments produce an intermediate output, G, with a standard production function:

$$G = f(L, K, Z) \quad , \tag{1}$$

where L is labor, K is capital equipment or facilities, and Z is other factor inputs. Assuming cost minimization, the associated first-stage cost function is:

$$TC = c(G, W) \quad , \tag{2}$$

where W represents factor prices.

Although factor inputs and government outputs can be separated conceptually, the difficulties encountered in measuring government outputs have led many researchers to use proxy variables for G. These proxies generally fall into one of two classes. Some are bundles of factor inputs, such a equivalent teacher hours or fire-fighting companies. These measures ignore variation in input mix and therefore implicitly assume that the production function in equation (1) exhibits Leontief technology. Others are measures of service outcomes, such as arrests, that are not

independent of the environment in which they are provided, so that they blur the distinction between the intermediate and final government outputs.

Government activities, G, are an intermediate output in the production of the final output or service outcome of interest to voters, S. The distinction between G and S is important because exogenous "environmental" factors influence the transformation of G into S. Two communities of the same size may utilize the same technology and level of resources for fire protection but experience significant differences in property losses and casualties, due to differences in the harshness of their fire-fighting environment.

Following Bradford, Malt, and Oates (1969), the second stage of the production process for S can be represented as

$$S = h(G, N, E) \quad , \tag{3}$$

where N is the jurisdiction's population and E represents environmental cost factors. Equation (3) indicates the level of governmental activity, G, required to produce a given level of public services, S, taking into account the impact of population, N, and the environment, E. Note that equation (3) can be solved for G, or

$$G = h^{-1}(S, N, E) \quad . \tag{4}$$

We will use this equation to develop public cost functions.

Population, N, is included in the final-output function to allow for the possibility of nonrivalry in the provision of public services. Nonrivalry exists if a 1 percent increase in population requires less than a 1 percent increase in the intermediate output, G, to maintain the

same level of final output, S.[2] Equation (4) reveals that population plays the same role as, and can be treated as, an environmental factor.

The cost of providing the final output depends not only on the production technology and factor prices, but also on the influence of environmental factors. Substituting equation (4) into equation (2) results in a second-stage cost function of the form:

$$TC = \hat{c}[h^{-1}(S, N, E), W] \quad . \tag{5}$$

Following Bradford, Malt, and Oates (1969), this derivation includes environmental factors in the second stage of the production function, equation (4), but not in the first stage, equation (2). In other words, equation (5) is based on the assumption that environmental factors do not affect the shape of the isoquants and hence do not affect factor substitution. This assumption may not be warranted; in the case of fire protection, for example, population density or building height may affect the trade-off between fire trucks, stations, and firefighters in producing fire-fighting companies. Our approach, therefore is to use a more general form of the cost function that does not require this assumption, and to determine empirically whether the assumption holds, namely:

$$TC = c[h^{-1}(S, N, E), W, E] \quad . \tag{6}$$

Equation (6) provides a flexible form for bringing final outputs, population, environmental factors, and factor prices into an analysis of public production. This approach has several

[2]An equivalent approach, used by Brueckner (1981) and Craig (1987), is to say that nonrivalry exists if a 1 percent increase in population leads to a less than 1 percent decrease in S, holding G constant.

advantages. It is based on final outputs, for which some measures exist. It requires a minimum of restrictions on the production technology, and it clearly separates the two stages of the production process. As we show in the next section, this approach also leads directly to clear definitions of returns to scale.

Defining Returns to Scale

As noted earlier, returns to scale are revealed in the relationship between average cost and the level of output. This relationship is more complicated in the public than in the private sector, however, because scale has two dimensions in the public sector, namely the quality of the public service and the number of people served.

Let us first consider the quality dimension. Service quality is measured by the final government output, S, and average cost, that is, costs per unit of quality, is measured by TC/S. We define "returns to quality scale" as the change in TC/S which results from a change S holding the service population, N, and environmental factors, E, constant. Using equation (6), this derivative is:

$$\frac{\partial\left(\frac{TC}{S}\right)}{\partial S} = \frac{\frac{\partial TC}{\partial G}\frac{\partial G}{\partial S} - \frac{TC}{S}}{S} = \frac{MC_S - AC_S}{S} , \tag{7}$$

where MC_S and AC_S are the marginal and average costs of producing S. Thus, increasing returns to quality scale exist if $MC_S < AC_S$, that is, if the average cost curve is downward sloping. Equation (7) can also be expressed in elasticity form:

$$\frac{\partial\left(\frac{TC}{S}\right)}{\partial S}\,\frac{S}{\frac{TC}{S}} = \left(\frac{\partial TC}{\partial G}\,\frac{G}{TC}\right)\left(\frac{\partial G}{\partial S}\,\frac{S}{G}\right) - 1 = \Theta_1\Theta_2 - 1 \quad . \tag{8}$$

Economies to quality scale exist if this expression is negative, that is if the product of Θ_1 and Θ_2 is less than unity.

The first of these elasticities, Θ_1, which we call "technical returns to scale," is the notion most closely associated with the definition of returns to scale in private production. It represents the technical relationship between inputs and the intermediate output of government. This interpretation is possible because of the duality between production and cost functions; Θ_1 equals the inverse of the elasticity of output, G, with respect to the scale of the inputs.[3] Increasing technical returns to scale are generally associated with production processes that involve extensive capital investment, so that the fixed cost of capital is large relative to the marginal cost of increased output, resulting in a declining average cost curve.

The second elasticity, Θ_2, measures what we call the "environmental effect," that is, the influence of the environment on the relationship between the intermediate and final output of government. Communities with a harsh environment require more G to obtain a given S than

[3]This result is proved by Varian (1984, pp. 68-9). It should be pointed out that changing input scale, that is, increasing all inputs by the same proportion, is not the same as increasing output while maintaining the cost-minimizing input mix, unless the scale change is very small or the production function is homogeneous with respect to all inputs.

do communities with a favorable environment. They are less likely, therefore, to face increasing returns to quality scale, even if all communities have the same technical returns to scale.[4]

Now let us turn to the population dimension, on which most of the research on public returns to scale has focused. This dimension involves the relationship between the per capita cost of the final government service, TC/N, and the number of people served, N. We define "returns to population scale" as the derivative of TC/N with respect to N, holding S and E constant. Again using equation (6), this derivative is:

$$\frac{\partial\left(\frac{TC}{N}\right)}{\partial N} = \frac{\frac{\partial TC}{\partial G}\,\frac{\partial G}{\partial N} - \frac{TC}{N}}{N} = \frac{MC_N - AC_N}{N} \tag{9}$$

Economies (diseconomies) to population scale exist if the per capita cost curve is downward (upward) sloping. This result also can be expressed in elasticity form:

$$\frac{\partial\left(\frac{TC}{N}\right)}{\partial N}\,\frac{N}{\frac{TC}{N}} = \left(\frac{\partial TC}{\partial G}\,\frac{G}{TC}\right)\left(\frac{\partial G}{\partial N}\,\frac{N}{G}\right) - 1 = \Theta_1\Theta_3 - 1 \tag{10}$$

Economies (diseconomies) to population scale exist if this expression is less than (greater than) zero.

As with returns to quality scale, the first elasticity, Θ_1, is technical returns to scale. In this case, however, the other elasticity, Θ_3 is the "nonrivalry parameter." It measures how much

---

[4]In the case of fire suppression, the key element for reducing property loss is speed, both the speed at which the fire is spreading and the response time of the fire department. The major environmental factor affecting the speed of fire spread is building condition. See Schaenman and Swartz (1974). Thus, the fire-fighting activities provided by a fire department interact with building condition to affect the cost of a given level of final output.

additional G is needed to maintain a given level of S when the number of people served increases and the effect of environmental factors is held constant. The value of Θ_3 is one for a private good and zero for a pure public good.

Note that two public services may face the same technical returns to scale but have different returns to population scale because of differences in their degree of nonrivalry. For example, police protection and garbage collection are often hypothesized to exhibit constant technical returns to scale. however, there may be some nonrivalry in the case of police protection, at least for small to medium-sized jurisdictions, whereas garbage collection is usually considered to be a private good.[5] Thus equation (10) indicates that, over some population range, police services may encounter increasing returns to population scale, whereas garbage collection is likely to face constant returns to population scale. In addition, a jurisdiction could face diseconomies of population scale even for a good that exhibits nonrivalry, because a high value for Θ_1 may offset a value of Θ_3 below one.

3. A Review of the Literature on Public Returns to Scale

Returns to scale in public production have been extensively studied, but the methodologies and results in this research vary significantly. Using the above conceptual framework, we sort the existing studies into three classes, according to the three concepts of returns to scale: technical returns to scale, returns to population scale, and returns to quality scale.[6]

---

[5]The results of Craig (1987) support the existence of nonrivalry in the case of police services, but the results of Ladd and Yinger (1989), which apply to very large cities, suggest severe congestion for police services, and hence diseconomies to population scale.

[6]See Duncombe (1989), Chapter 2, for a more detailed review of the literature on economies of scale.

Technical Returns to Scale

Several studies have examined the relationship between total costs and some measure of the intermediate government output, G. The earliest of these studies use a quadratic form to allow for a "U-shaped" average cost curve (see Hirsch, 1965). Some recent studies have employed more flexible cost functions, which place few restrictions on the characteristics of the production technology (see Deller et al., 1988, and Gyapong and Gyimah-Brempong, 1988). The results of these studies vary significantly, due in part to differences both in the way the intermediate output is measured and in the sets of environmental factors used as controls.

The measures of output usually fall somewhere between an intermediate output and a final output, as we have defined the terms. For example, the number of arrests clearly depends on the extent of criminal activity, an environmental factor. As a result, technical economies cannot be disentangled from the effect of environmental factors on costs without carefully controlling for environmental factors. Although a few of these studies include environmental variables in their regressions, none of them explicitly separates these two aspects of public production. The approach we present below, which is based on equation (6), allows us to make this separation.

Returns to Population Scale

Much of the early research on returns to scale in the public sector focused on the relationship between population and public expenditure.[7] To be specific, two groups of studies have addressed this relationship.

The first group consists of studies that aim for direct evidence on returns to population scale. The best of these studies provide cross-section regressions of per capita expenditure against a

---

[7]To some degree this focus reflects the close link between population and popular policy issues, such as optimal city size and metropolitan consolidation. It also reflects the importance of population in determining which public services are Samuelsonian public goods.

jurisdiction's population, holding service quality and other cost factors constant (Ahlbrandt, 1973). Because these studies are not based on a formal analysis of public service costs, however, they cannot separate the two components of returns to population scale, namely technical returns to scale and nonrivalry.

The second group addresses the relationship between expenditure and population in the context of a median-voter model of local spending. Beginning with Borcherding and Deacon (1972), the studies in this group have focused on the concept of nonrivalry.[8] This concept is introduced through a "congestion function" of the form: $S = (G)(N^{-g})$, where $g = 0$ for a pure public good and $g = 1$ for a private good. The congestion parameter, g, conveys the same information as our nonrivalry parameter, Θ_3. This function implies that population influences the amount the median voter must pay for another unit of S, that is, her tax price, and therefore influences the level of S she selects.

This congestion function is a special case of the final-output function, equation (3). As noted earlier, population can be thought of as an example of an environmental variable, in that it does not affect technical returns to scale, but does affect the relationship between G and S. Models of local spending usually find congestion parameters of one or greater. This research can be criticized on several ground. From our standpoint, the most important criticism is that these studies do not distinguish between congestion and returns to population scale.[9] Moreover, as

---

[8]See Inman (1979) and Rubinfeld (1985) for reviews of this literature.

[9]Romer and Rosenthal (1979) first raised this issue but did not provide any further explanation. These studies also have been criticized for not holding the range of services provided by a government constant (Oates, 1988 and McMillan, 1989) and for using a restrictive form for the congestion function (Craig, 1987, and Edwards, 1990).

shown below, the congestion parameter cannot be identified in a median-voter model without assuming constant returns to quality scale.

Several recent articles have gone beyond the estimation of a congestion parameter. Brueckner (1981) sets up a function similar to our equation (3), and measures congestion by the elasticity of S with respect to N.[10] Brueckner also recognizes that one must distinguish between congestion and technical returns to scale, Θ_1, although he assumes that technical returns to scale are constant. He also estimates his function for fire protection and finds, contrary to previous research, significant nonrivalry in the provision of fire protection.

In addition, Brueckner proposes a summary notion of "returns to scale in consumption" for local public goods.[11] According to this notion, a local public service exhibits increasing returns to scale if the final output increases when both population and the intermediate good increase by the same proportion, that is, when population increases and the per capita amount of the intermediate good is held constant. In fact, however, Brueckner's proposed measure just rescales the congestion elasticity.[12]

---

[10]Brueckner's approach differs from ours in that his intermediate output is public good capacity instead of the level of governmental activity. We do not regard this as a significant difference.

[11]We find this terminology to be confusing. It implies that returns to scale are in some way a demand phenomenon, whereas they are, in both Brueckner's model and ours, strictly an aspect of public production.

[12]Totally differentiating equation (4) with respect to G, S, and N and imposing Brueckner's condition that $dG/G = dN/N$, we find that his returns to scale measure, in elasticity form, equals $(1 - \Theta_3)/\Theta_2$. Because Θ_2 is always positive, the sign of this expression, which determines whether there are economies or diseconomies of scale, depends only on Θ_3. It should also be pointed out that Brueckner's approach breaks the duality between production and cost measures. To be specific, equations (4) and (6) imply that when Brueckner's condition is met, the elasticity of TC/N with respect to N equals $(\Theta_1 - 1)$. When the ratio of G to N is held constant, per capita costs depend only on technical returns to scale. Given Brueckner's assumption about Θ_1, this elasticity always equals zero.

Craig (1987) also begins with equation (3), and he estimates separate production functions for the intermediate and final outputs of police services. To facilitate the estimation of congestion, he uses a simple production function for the intermediate output, G, which he defines to be clearances or crimes solved by the police.[13] This measure is not independent of the environment, so his production function for G includes not only factor inputs, but environmental factors and population. Because he employs a measure of G that is affected by environmental factors and a restrictive form for its production function, Craig is unable to estimate technical returns to scale or to isolate the role of the environment.

Returns to Quality Scale

Although they do not formally identify the concept, Baum (1986) and Gyimah-Brempong (1989) provide the first estimates of returns to quality scale. Using measures of a final output, they find diseconomies to quality scale in the case of education and policy, respectively. Baum's cost model includes only current expenditure and the price of labor. Because technical economies of scale are usually associated with the existence of capital facilities or equipment, the exclusion of capital from his model may explain his result. Gyimah-Brempong (1989) does not control for population in his cost model, so that his result is subject to left-out-variable bias. Moreover, neither cost model can separate the influence of technology and of environmental factors on returns to quality scale.

---

[13]Craig (1987) assumes that other inputs, such as capital, are constant, so that police officers are the only variable factor input. He then specifies a general production function, which supposedly allows nonconstant returns to scale. With only one variable input, however, he appears to be forcing decreasing technical returns to scale. His measure of final output is an estimate of the actual (as opposed to reported) crime rate.

4. Estimating Returns to Scale in Public Production

In this section we present an empirical cost model for estimating the components of returns to scale in the public sector. Much of the research on local public production focuses on public spending, not public outputs. This approach has the great advantage that it can make use of high-quality, readily available expenditure data. However, technical returns to scale cannot be estimated from an expenditure model. After demonstrating why this is true, we present an alternative approach, based on public outputs, that makes it possible to estimate all aspect of public returns to scale.

The Limits of Expenditure Models

All of the studies in the large literature on local public spending assume, explicitly or implicitly, constant returns to quality scale. This assumption is necessary to make expenditure models tractable (see Inman, 1979).

To illustrate the problem, let us assume that the intermediate output, G, is produced using Cobb-Douglas technology with two factors of production, labor, L, and capital, K, and that local governments minimize costs subject to a production constraint. These assumptions imply that the first-stage cost function is of the form;

$$TC = A\, G^{\frac{1}{\kappa}}\, w^{\frac{\alpha}{\kappa}}\, r^{\frac{\beta}{\kappa}} \quad , \tag{11}$$

where A is a constant, w and r are the prices of labor and capital, α and β are output elasticities, and $\kappa = (\alpha+\beta)$ measures technical returns to scale.

let us further assume that the final output, S, is produced using a constant elasticity function of the intermediate output, G, population, N, and environmental factors, E; that is,

$$S = B\ G^{\lambda}\ N^{-g}\ E^{\nu} \quad , \tag{12}$$

or

$$G = B^{-\frac{1}{\lambda}}\ S^{\frac{1}{\lambda}}\ N^{\frac{g}{\lambda}}\ E^{-\frac{\nu}{\lambda}} \tag{13}$$

Note that B is a constant and, as before, g is the congestion parameter. Moreover, λ could be specified as a function of E to add another dimension to the impact of environmental factors on the translation of G into S. Substituting equation (13) into equation (11) results in the following second-stage cost function, where A' is another constant:

$$TC = A'\ S^{\frac{1}{\lambda\kappa}}\ N^{\frac{g}{\lambda\kappa}}\ E^{\frac{\nu}{\lambda\kappa}}\ w^{\frac{\alpha}{\kappa}}\ r^{\frac{\beta}{\kappa}} \tag{14}$$

To account for the fact that total cost (or expenditure) and service quality are set simultaneously in the local political process, many studies combine equation (14) with a model of public choice. The most common approach is to use a median voter model with a constant elasticity demand function, or:[14]

---

[14]See Bergstrom and Goodman (1973) for the conditions under which the median voter model holds. See Inman (1979) and Rubinfeld (1985) for surveys of the literature. To simplify the presentation, a few demand variables included in the estimation are left out of equation (15).

$$S = C\, Y_m^{\psi} \left(\frac{V_m}{\bar{V}}\right)^{\mu} MC^{\mu} \quad , \tag{15}$$

where C is a constant, Y_m is median income, $V_m/\bar{V}$ is the ratio or median to mean house value or the median voter's tax share, MC is the marginal cost of producing S, ψ is the income elasticity, and μ is the price elasticity of demand for S.

To derive an expenditure function, we find the marginal cost function from (14) and substitute it into (15). Solving for S and substituting the result back into (14) yields:

$$TC = A''\, Y_m^{\frac{\psi}{\sigma}} \left(\frac{V_m}{\bar{V}}\right)^{\frac{\mu}{\sigma}} N^{\frac{g(\mu+1)}{\sigma}}\, E^{\frac{\nu(\mu+1)}{\sigma}}\, r^{\frac{\lambda\beta(\mu+1)}{\sigma}}\, w^{\frac{\lambda\alpha(\mu+1)}{\sigma}} \tag{16}$$

where $\sigma = [\lambda\kappa(1 + \mu) - \mu]$ and A'' is a constant.

Equation (16) is log-linear so it can be estimated with linear regression techniques. Excluding the constant term, it includes six coefficients containing seven parameters (ψ, μ, g, ν, λ, α, β). As a result, the parameters are not identified. Even if one has a priori knowledge about λ, the environmental elasticity, it is not possible to identify technical returns to scale, κ, which are part of σ, the denominator of every coefficient.[15] The only way out of this problem is to assume a value for κ. All the studies in this literature, except for Baum (1967) and Gyimah Brempong (1989), assume $\kappa = 1$, that is, constant technical returns to scale.

[15]In more general terms, the parameters are not identified with λ known because the second coefficient is a linear combination of the fifth and sixth. See Duncombe (1989) for a proof.

Although we have not provided a general proof that technical returns to scale cannot be estimated in an expenditure model, production functions other than the Cobb-Douglas are much harder to work with and appear to make identification even more difficult. The general problem is that because expenditure data do not allow the separation of average cost and quantity, they do not provide sufficient information to identify the impact of changes in scale on average cost. Estimating technical returns to scale requires additional information about the final output, intermediate output, or factor inputs.

A Cost Model for Estimating Public Returns to Scale

In this section we specify a cost model with which technical returns to scale, as well as the environmental effect and nonrivalry, can be estimated. This model requires a measure of the final output, S, and a specific form for the final-output function, equation (3). In the following section we estimate this model using data on fire departments in New York State.

We assume that the final-output function takes the form given by equation (12) and that λ, the elasticity of S with respect to G, depends on a subset of the environmental factors, E*.[16] To be specific, we assume that $1/\lambda = (1 + \lambda^* E^*)$. This specification is required to allow separate estimation of the environmental effect, $1/\lambda$, and technical returns to scale, Θ_1 (see Duncombe, 1989).

---

[16]This form of the final output function assumes decreasing marginal congestion, or $\partial^2 S/\partial N^2 > 0$. In other words, the marginal impact of population on the final output goes down as population increases. Using New York expenditure data, Edwards (1990) estimates this function as well as a more general function which did not impose restrictions on the sign of $\partial^2 S/\partial N^2$. For two of the three services examined (police and general administration) he was not able to find a statistically significant difference between the models. Note, however, that Edwards assumes constant returns to quality scale.

A Cobb-Douglas technology, which leads to the cost function in equation (14), restricts the elasticity of factor substitution between any pair of factors to be unity, and implies that technical returns to scale does not vary with the output level. As a result, we employ a more flexible cost function, namely the "translog" function developed by Christensen, Jorgenson and Lau (1973). One by-product of this approach is that we can test whether a Cobb-Douglas technology fits fire service production.

We assume that the cost of G can be described by a translog cost function for N factors of production. We modify the standard translog function by substituting equation (13) for G. In a translog model, factor substitution is estimated from the coefficients of the factor-price terms. To allow for the potential impact of some set of environmental factors, E', on factor substitution, therefore, we multiply all the factor price terms by $(1 + \lambda' E')$. This procedure allows us to test whether $\lambda' = 0$, that is, whether equation (6) can be simplified to equation (5). This modified translog function, with M environmental factors, can be represented by:

$$\ln TC = A + \delta_1 \ln G + \frac{1}{2}\delta_2 (\ln G)^2 + (1 + \lambda' E') \left[\sum_i^N \varepsilon_i \ln W_i + (\ln G) \sum_i^N \psi_i \ln W_i + \frac{1}{2} \sum_i^N \sum_j^N \gamma_{ij} (\ln W_i)(\ln W_j) \right], \tag{17}$$

where W_i is the price of factor i and, from equation (13),

$$\ln G = (1 + \lambda^* E^*)[\ln S + g \ln N - \sum_k^M \nu_k \ln E_k] .$$

A factor-share equation is usually estimated along with a translog cost function to add information. Assuming cost minimization, Sheppard's lemma implies that the derivative of lnTC with respect to $\ln W_i$ is equal to the share of total costs spent on factor i, namely Sh_i. Using equation (17), we find that the factor share equations are:

$$Sh_i = (1 + \lambda' E')\left[\varepsilon_i + \psi_i \ln G + \sum_j^N \gamma_{ij} \ln W_j\right] . \tag{18}$$

for i = 1 to (N-1).

In summary, the translog cost model for public services involves the estimation of N equations--equation (17) and (N-1) factor share equations, (18)--for (N+1) endogenous variables: S, TC, and Sh_i.[17] Our approach differs from previous translog studies because, as explained below, the final output, S, is treated as endogenous.

This model makes it possible to test hypotheses about the form of the production technology. For example, it can be used to test for Cobb-Douglas technology or the homogeneity of the cost function with respect to the final output.[18] Rejection of the Cobb-Douglas restriction but not

---

[17]To avoid singularity of the covariance matrix, only (N-1) factor share equations are estimated.

[18]See Brown, Caves and Christensen (1979) for presentation of the various restrictions. Any cost function must exhibit linear homogeneity in factor prices, which using equation (17) implies the following restrictions;

$$\sum_i^N \varepsilon_i = 1 \qquad \sum_i^N \psi_i = 0 \qquad \sum_i^N \gamma_{ij} = 0 \ .$$

and $\gamma_{ij} = \gamma_{ji}$ for all i, j. The additional restrictions for homogeneity in output include $\delta_2 = 0$ *and* $\psi_i = 0$; and for Cobb-Douglas technology, $\delta_2 = 0$ and ψ_i, $\gamma_{ij} = 0$, for all i,j.

the homogeneity restriction would imply that a more general homogeneous function, such as the CES, is appropriate.

The translog model also makes it possible to estimate both returns to quality scale and to population scale. using equation (17), we find that returns to quality scale equal:

$$\frac{\partial \ln TC}{\partial \ln S} = \Theta_1\Theta_2 = (1+\lambda^* E^*)\ [\delta_1 + \delta_2(\ln S + g\ \ln N - \sum_k^N \nu_k \ln E_k) + (1+\lambda' E')\sum_i^N \psi_i \ln W_i\ ] \quad , \tag{19}$$

where $\Theta_2 = (1+\lambda^* E^*)$ and Θ_1 is the term in square brackets. Similarly, returns to population scale from equation (17) equal:

$$\frac{\partial \ln TC}{\partial \ln N} = \Theta_1\Theta_3 - g(1+\lambda^* E^*)\ [\delta_1 + \delta_2(\ln S + g\ \ln N - \sum_k^N \nu_k \ln E_k) + (1+\lambda' E')\sum_i^N \psi_i \ln W_i\ ] \tag{20}$$

where $\Theta_3 = g(1+\lambda^* E^*)$. Remember that for both types of scale, derivatives less than (greater than) unity indicate increasing (decreasing) returns to scale.

5. Estimates of Returns to Scale in Fire Protection

We estimate our cost model for 197 paid, municipal fire departments in New York State in 1984-86.[19] This section describes our measures, data sources, and results.

Measures and Data Sources

The first data issue is the definition and measurement of fire service output. The most direct measure of the final output from fire services is the percentage of a community property value that is lost to fire during a year. The source of property loss data is the New York Office of Fire Prevention and Control (OFPC).[20]

The major inputs used in the production of fire services can be aggregated into two broad categories; labor and capital equipment plus facilities. The sources used for firefighter salaries are the "Labor Agreements Data System" (LADS) developed by the New York Department of State and the *1982 Census of Governments*. The appropriate price of capital is the annual rental price for a unit of capital services.[21] The rental price of capital has been modelled as: $r = q(d + i)$, where q is the purchase price of a capital asset, d is the depreciation rate and i is

---

[19]Volunteer fire departments are excluded because the recorded expenditure for personnel is generally zero, which understates the true opportunity cost of hiring volunteers. Because the exclusion of volunteer departments may bias the results estimated for the paid departments, we use the methodology developed by Heckman (1976) to control for this potential sample selection bias. See Duncombe (1989) for the details.

[20]The denominator of this measure is total full-market property value, estimated by the New York State Board of Equalization and Assessment.

[21]Several recent studies of public service costs use either the purchase price or the replacement cost as the price of capital. See Gyapong and Gyimah-Brempong (1988), Gyimah-Brempong (1989), and Deller, Chicoine, and Walzer (1988). These studies appear to have ignored the intertemporal nature of capital investment which, by definition, has a useful life over one year.

the real interest rate.[22] The other principal endogenous variable is the total cost of the resources used in the production of fire services. Because expenditure data measure when resources are legally committed not when they are used, we approximate capital cost using the average real capital expenditure over a five-year period. Our expenditure data were obtained from the New York State Comptroller.

In addition to population, N, environmental factors affecting fire service costs fall into three groups: building condition, the presence of high-risk property, and the probability that fires will spread. We measure building condition, E1, by the sum of the percentage of housing units built before 1940 and the percentage of families below the poverty line.[23] The presence of high-risk property is indicated by the percentage of the community's property value in industrial and utility property, E2, and the percentage of housing units in buildings higher than two stories, E3. Finally, the probability that fires will spread is measured by E3 and the logarithm of population density, E4. Because building condition is the environmental factor most likely to affect the elasticity of S with respect to G, E^* in equation (17) is assumed to be equivalent to E1.[24] The two environmental variables most likely to affect factor substitution are population density, E4, and building height, E3. On the basis of the appropriate statistical test, we find that building

---

[22]The rental price of capital is analyzed in the research on private sector investment (see Jorgenson, 1974). We construct a measure of bond yield for municipalities in New York, which is combined with information on purchase prices and depreciation rates to obtain a measure of the rental price of fire-department capital in New York State. See Duncombe (1989). .

[23]An alternative approach with building condition measured by the product of these two variables yields virtually identical results.

[24]Our regressions also include variables to control for service responsibilities (the percentage of a fire department's responses that involve emergency medical service, E5, and that involve mutual aid, E6, and for sample selection (the percentage of fire-fighters who are paid, E7, and Heckman's lambda, E8).

height does not affect factor substitution, and E' is assumed to be simply E4. The principal source for socio-economic data is the *1980 Census of Population and Housing*.

Empirical Results

We estimate the translog cost function, equation (17), and a labor share function, equation (18), using nonlinear 2SLS, with variables from a median voter demand model as instruments.[25] To test hypotheses about model specification, we employ the likelihood-ratio test for nonlinear models developed by Gallant (1987). The results of our specification tests are presented in Table 1.[26]

---

[25]The differences between the 2SLS and 3SLS estimates are not significant. All equations were estimated with SAS SYSNLIN, with data scaled appropriately to guarantee the accuracy of the results.

[26]Two tests are not in Table 1. First, inferring production characteristics from cost functions requires exogenous factor prices. We test for the endogeneity of labor and capital prices using the specification test developed by Hausman (1978) and cannot reject the hypothesis that they are exogenous. Second, we test the restriction for cost minimization, namely linear homogeneity in factor prices. Although this restriction can be rejected at the 90 percent level, we nevertheless assume that it is met to assure duality between production and cost functions. Developing a model in which public decision makers do not minimize costs is beyond the scope of this paper.

TABLE 1
HYPOTHESIS TESTS RELATED TO RETURNS TO SCALE[a]
(chi-square distribution)

	Degrees of Freedom	Critical[b] Value	Calculated Value
Environment does not Affect Returns to Quality Scale ($\lambda^* = 0$)	1	2.71	2.08
Environment does not Affect Factor Substitution ($\lambda = 0$)	1	2.71	12.64
Homogeneity in S ($\psi_1 = 0, \delta_2 = 0$) ($\delta_2 = 0$)	 2 1	 4.61 2.71	 87.22 4.68
Cobb-Douglas Technology ($\psi_1 = 0, \delta_2 = 0, \gamma_1 = 0$)	3	6.25	105.36

[a]Based on a likelihood ratio test for linear restrictions on a nonlinear simultaneous equations model (translog cost function). The test statistic has a chi-square distribution. See Gallant (1987) for methodology. Degrees of freedom are equal to the linear restrictions on the model.

[b]For statistical significance at the 90 percent level.

The first specification tests concern the environmental factors. First, we examine the null hypothesis that building condition does not affect the elasticity of S with respect to G, that is, that $\lambda^* = 0$. We cannot reject this null hypothesis even at the 90 percent level. This result implies that $\Theta_2 = 1$, so that, in elasticity form, technical returns to scale and returns to quality scale are equivalent in our sample. Second, we test the hypothesis that population density does not affect the shape of the isoquants, that is, $\lambda' = 0$. We can reject this null hypothesis at the 99 percent confidence level. In the case of fire protection, therefore, we reject the Bradford, Malt, and Oates (1969) assumption that environmental factors only affect the second-stage production function.

We then test the hypothesis that the cost function is homogeneous in the final output, S. This hypothesis can be rejected at the 95 percent level.[27] This result indicates that returns to quality scale vary with output so that neither the Cobb-Douglas nor the CES production functions are appropriate for fire services. Another specification test allows us to explicitly reject Cobb-Douglas technology. See Table 1.

The detailed empirical results from the final translog cost model for fire services are presented in Table 2. Based on these results, we calculate returns to quality scale and to population scale. At the mean level of quality, the elasticity of total costs with respect to G equals 0.21, indicating significant technical economies of scale. This result runs counter to the almost-universal assumption of constant technical returns to scale for fire protection. In addition, we find that the environmental elasticity is unity, which implies that, in elasticity form, returns to quality scale equal technical returns to scale minus 1, or negative 0.79. This result indicates increasing returns to quality scale.[28]

Table 2 also presents our results for returns to population scale. Our estimated nonrivalry elasticity is 5.18, which indicates far more congestion (i.e., less nonrivalry) in fire services than is found by previous studies, such as Brueckner (1981). Combining this result with our estimate

---

[27]We also tested the hypothesis that the coefficient of $(\ln G)^2$, δ_2, is zero. This hypothesis can be rejected at the 95 percent confidence level.

[28]Moreover, the estimated economies to quality scale are roughly the same at different levels of S or average costs. For example, returns to quality scale range from a low of -0.82 for fire districts in the highest quartiles of both S and AC to a high of -0.76 for the districts in the lowest quartiles.

TABLE 2

ESTIMATED COEFFICIENTS FOR THE TRANSLOG COST (AND LABOR SHARE) MODEL[a] FIRE PROTECTION IN NEW YORK STATE

Coefficient	Estimate	Coefficient	Estimate
A	-9.8694* (-4.89)	ν_1	-2.5989* (2.63)
δ_1	0.3660* (3.60)	ν_2	1.7833 (1.90)
δ_2	-0.0055** (-1.92)	ν_3	0.6608 (0.39)
γ'	0.067 (3.91)	ν_4	-3.8993* (-3.21)
ε_1	0.3222* (6.23)	ν_5	1.1411* (2.64)
ψ_1	0.0082* (4.10)	ν_6	-1.3269 (-0.65)
λ_1	0.0423 (1.41)	ν_7	2.5825* (4.26)
g	5.1847+ (3.95)	ν_8	-0.5145 (-1.53)
Economies of Quality Scale[b]			
Technical Economies of Scale ($\partial \ln TC/\partial \ln G = \Theta_1$)			0.21
Environmental Effect ($\partial \ln G/\partial \ln S = \Theta_2$)			1.00
Economies of Quality Scale ($\partial \ln(TC/S)/\partial \ln S = \Theta_1\Theta_2 - 1$)			-0.79
Economies of Population Scale[b]			
Technical Economies of Scale ($\partial \ln TC/\partial \ln G = \Theta_1$)			0.21
Congestion Elasticity ($\partial \ln G/\partial \ln N = \Theta_3$)			5.18
Economies of Population Scale ($\partial \ln(TC/N)/\partial \ln N = \Theta_1\Theta_3 - 1$)			0.09

*Significantly different from zero at the 95 percent level.

**Significantly different from zero at the 90 percent level.

+Significantly different from one at the 95 percent level.

[a]Estimated with nonlinear 2SLS' t-statistics are in parentheses; 197 observations. Refer to equations (17) and (18) for coefficients where W_1 is the price of labor and W_2 is the price of capital. The full restrictions for linear homogeneity with respect to factor prices are imposed.

[b]Calculated at the mean value for all variables.

of technical returns to scale, we find that the elasticity of per capita costs with respect to population is between 0.0 and 0.1, indicating constant returns to population scale.[29]

6. Conclusions

In the public sector, returns to scale have two dimensions, quality and population, which are not clearly distinguished in previous studies. We show how to estimate both dimensions without imposing restrictions on public technology. We implement our approach, which depends on the availability of a measure of the final government output, for paid fire departments in New York State.

We find that the elasticity of total costs with respect to the intermediate output is less than unity, indicating increasing technical returns to scale. Because our estimated environmental elasticity is one, this result also implies increasing returns to quality scale. We also find that the nonrivalry elasticity is greater than unity. Somewhat surprisingly, increasing technical returns to scale and this nonrivalry result exactly offset each other, implying constant returns to population scale.

Most previous studies of public production have assumed constant technical returns to scale. Our results strongly contradict this assumption. Because returns to quality scale and returns to population scale both depend on technical returns to scale, our results suggest that previous studies of public production may be misspecified. In addition, many previous studies assume that public production functions are homogeneous, and those studies that include environmental variables follow Bradford, Malt, and Oates (1969) in ignoring the impact of these variables on

[29]Returns to population scale are roughly constant at all values of N and (TC/N). They range from a low of -0.07 for fire districts in the lowest quartiles of N and (TC/N) to a high of 0.20 for districts in the highest quartiles.

factor substitution. We can reject the hypothesis that the production function is homogeneous, but not the hypothesis that environmental variables affect factor substitution. These results cast further doubt on the specifications in previous studies.

Our results may also be of interest to policymakers. To the extent that decisions about fire service quality are based on the view that the cost per unit of quality is constant, fire officials (or voters) are selecting fire protection levels that are too low in the sense that the marginal benefit of increments to quality would exceed the marginal cost. Moreover, constant returns to population scale imply that consolidating small fire departments will not result in cost savings.

References

Ahlbrandt, Jr., Rogers. 1973. "Efficiency in the Provision of Fire Services." *Public Choice* 16 (Fall):1-15.

Baum, Donald. 1986. "A Simultaneous Equations Model of the Demand for and Production of Local Public Services: The Case of Education." *Public Finance Quarterly* 14 (April):157-178.

Bergstrom, Theodore and Goodman Robert. 1973. "Private Demand for Public Goods." *The American Economic Review* 63 (June):286-296.

Borcherding, Thomas, and Deacon, Robert. 1972. "The Demand for Services of Non-Federal Governments." *The American Economic Review* 62 (December):891-901.

Bradford, David; Malt, Robert; Oates, Wallace. 1969. "The Rising Cost of Local Public Services: Some Evidence and Reflections." *National Tax Journal (June):185-202.*

Brown, Randall; Caves, Douglas; Christensen, Laurits. 1979. "Modelling the Structure of Cost and Production for Multiproduct Firms." *Southern Economic Journal* 46 (July):256-273.

Brueckner, Jan. 1981. "Congested Public Goods: The Case of Fire Protection." *Journal of Public Economics* 15 (February):45-58.

Craig, Steven. 1987. "The Impact of Congestion on Local Public Good Production." *Journal of Public Economics* 32:331-353.

Christensen, Laurits; Jorgenson, Dale; Lau, Lawrence. 1973. "Transcendental Logarithmic Production Frontiers." *The Review of Economics and Statistics* 55 (February):28-45.

Deller, Steven; Chicoine, David; Walzer, Norman. 1988. "Economies of Size and Scope in Rural Low-Volume Roads." *The Review of Economics and Statistics* 70 (August):459-465.

Duncombe, William. 1989. *Evaluation of Factors Affecting the Cost of Public Services With an Application for Fire Protection.* Doctoral dissertation (Syracuse, NY: Syracuse University).

Edwards, John. 1990. "Congestion Function Specification and the 'Publicness' of Local Public Goods." *Journal of Urban Economics* 27 (January):80-96.

Gallant, Ronald. 1987. *Nonlinear Statistical Models.* New York: John Wiley & Sons.

Gyapong, Anthony and Gyimah-Brempong, Kwabena. 1988. "Factor Substitution, Price Elasticity of Factor Demand and Returns to Scale in Police Production: Evidence from Michigan." *Southern Economic Journal* 54 (April):863-878.

Gyimah-Brempong, Kwabena. 1989. "Production of Public Safety: Are Socioeconomic Characteristics of Local Communities Important Factors?" *Journal of Applied Econometrics* 4 (January-March):57-71.

Hausman, Jerry. 1978. "Specification Tests in Econometrics." *Econometrica* 46 (November):1251-1271.

Heckman, James. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." *Annuals of Economic and Social Measurement* 5:475-492.

Hirsch, Werner. 1965. "Cost Functions of an Urban Government Service: Refuse Collection." *Review of Economics and Statistics* 47 (February):87-93.

Inman, Robert. 1979. "The Fiscal Performance of Local Governments: An Interpretative Review." In P. Mieszkowski and M. Straszheim (ed.) *Current Issues in Urban Economics* (Baltimore: The Johns Hopkins University Press), pp. 270-321.

Jorgenson, Dale. 1974. "Investment and Production: A Review." In M. Intrilligator and D. Kendrick (ed.), *Frontiers of Quantitative Economics* (Amsterdam: North-Holland Publishing Company), pp. 341-375.

Ladd, Helen and Yinger, John. 1989. *America's Ailing Cities: Fiscal Health and the Design of Urban Policy*. Baltimore: The John Hopkins University Press.

McMillan, Melville. 1989. "On Measuring Congestion of Local Public Goods." *Journal of Urban Economics* 26 (September 1989):131-137.

Oates, Wallace. 1988. "On the Measurement of Congestion in the Provision of Local Public Goods." *Journal of Urban Economics* 24 (July):85-94.

Romer, Thomas and Rosenthal, Howard. 1979. "The Elusive Median Voter." *Journal of Public Economics* 12:143-170.

Rubinfeld, Daniel. 1985. "The Economics of the Local Public Sector." In A. Auerback and M. Feldstein (ed.), *Handbook of Public Economics*, Vol. 2 (New York: Elsevier Science Publishers), pp. 571-645.

Schaenman, Philip and Swartz, Joe. 1974. *Measuring Fire Protection Productivity in Local Government*. Boston: National Fire Protection Association.