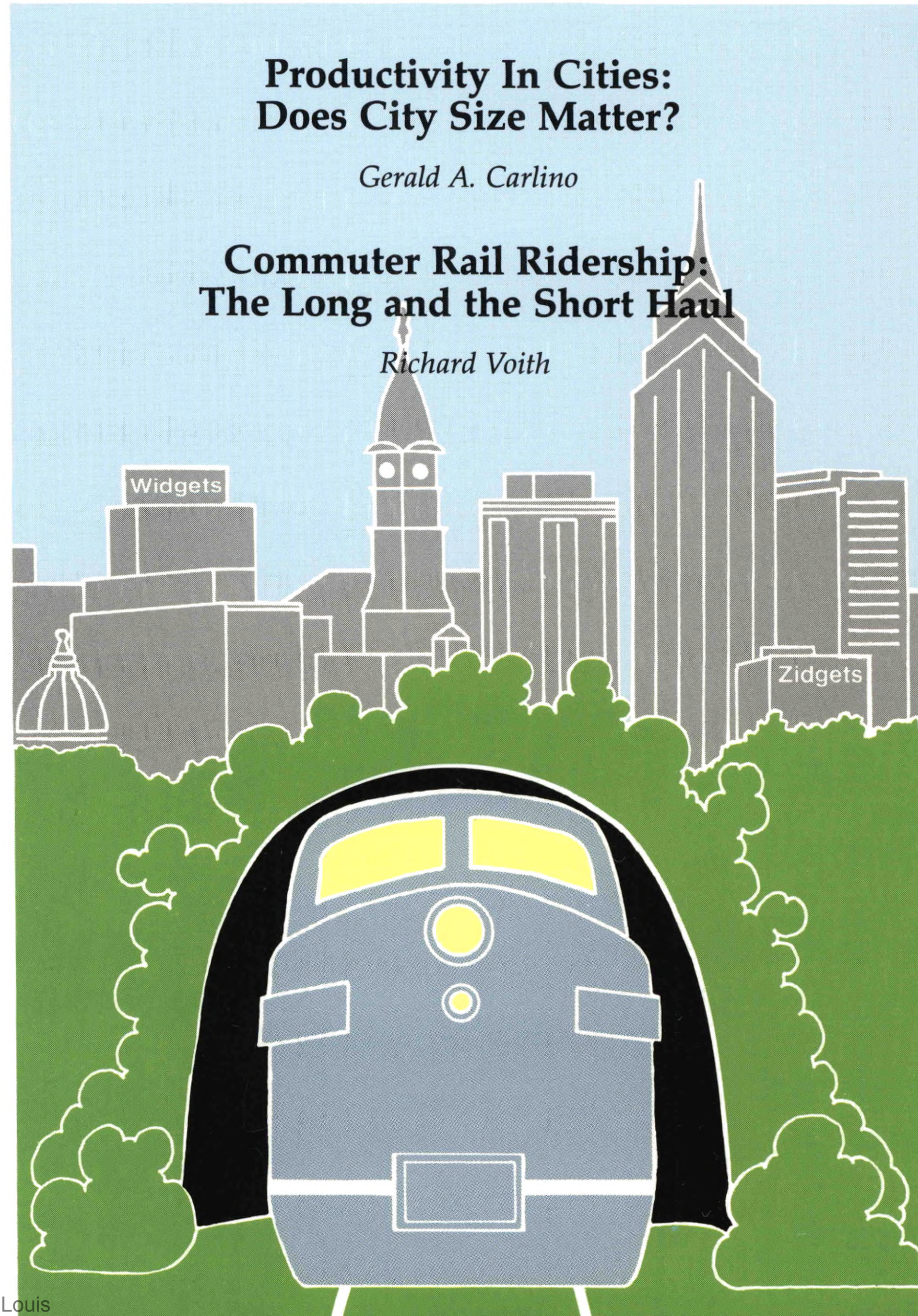
BUSINESS REVIEW
ISSN 0007-7011
Federal Reserve Bank of Philadelphia
NOVEMBER • DECEMBER 1987



BUSINESS REVIEW

The BUSINESS REVIEW is published by the Department of Research six times a year. It is edited by Judith Farnbach. Artwork is designed and produced by Dianne Hallowell under the direction of Ronald B. Williams. The views expressed herein are not necessarily those of this Reserve Bank or of the Federal Reserve System.

SUBSCRIPTIONS. *Single-copy subscriptions for individuals are available without charge. Institutional subscribers may order up to 5 copies.*

BACK ISSUES. *Back issues are available free of charge, but quantities are limited: educators may order up to 50 copies by submitting requests on institutional letterhead; other orders are limited to 1 copy per request. Microform copies are available for purchase from University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan, 48106.*

Please send subscription orders, back orders, changes of address, and requests to reprint to the Federal Reserve Bank of Philadelphia, Department of Research, Publications Desk, Ten Independence Mall, Philadelphia, PA 19106-1574, or telephone (215) 574-6428. Please direct editorial communications to the same address, or telephone (215) 574-3805.

NOVEMBER/DECEMBER 1987

Productivity in Cities: Does City Size Matter?

*Gerald A. Carlino**

INTRODUCTION

Economists have long recognized that a firm's size can affect its productivity. As a firm increases its size, it can sometimes increase productivity by having its workers specialize in particular tasks, or by using its capital equipment more efficiently. In these situations a firm is said to enjoy *internal* economies of scale.

Another important source of a firm's productivity that is often overlooked is a type of economies of scale that is *external* to the firm. These external economies of scale are also referred to as *agglomeration economies.* Increases in productivity due to agglomeration economies depend not upon the size of the firm itself, as in the case of internal economies of scale, but either upon the size of a firm's industry in a particular city or upon the size of the city itself.

To a large extent, market forces will encourage private firms to seek out and take advantage of agglomeration economies as they attempt to become more productive. But city planners also have a crucial role to play in accommodating such growth. If planners fail to address and

*Gerald A. Carlino is a Senior Economist and Research Advisor in the Urban and Regional Section of the Research Department at the Federal Reserve Bank of Philadelphia.

resolve the problems of congestion that arise as a city grows, firms will find their costs of doing business in that city increasing. The productivity gains from city size may not be fully tapped if the city cannot accommodate the growth which agglomeration economies spur. This means investing in public infrastructure, such as roads, bridges, sewers, and public transportation systems.

In addition to enabling firms to take advantage of agglomeration economies by investing in public infrastructure, local governments accommodate economic activity by making an investment in the people who live there. It is through education and training, which is primarily the responsibility of local authorities, that worker productivity is increased and that advances in technology are introduced to the labor force.

Economists have not only developed theories about why firms may be more productive in cities, but they have also attempted to measure how much agglomeration economies matter, and how much public infrastructure as well as education and training may influence productivity. Without exception, the effects of each of these factors on productivity have been found to be significant.

BIGGER IS OFTEN BETTER: WHY AGGLOMERATION ECONOMIES MEAN GREATER PRODUCTIVITY

Economists describe the advantages and disadvantages of a firm's expanding in terms of "returns to scale." Suppose a firm doubles all of its inputs in production, using twice as much raw material, twice as many workers, and twice as much capital equipment. If it more than doubles its original level of output, the firm is said to be enjoying increasing returns to scale, or *internal* economies of scale. In this case, bigger is better. If the firm doubles all of its inputs and produces exactly twice as much output as originally, economists refer to this as constant returns to scale. When a firm doubles all of its inputs and finds that its output is less than twice the original level, it has reached the point of decreasing returns to scale (diseconomies of scale). This typically occurs as the scale of a firm expands beyond a certain point, because management becomes less efficient in very large-scale operations. In this case, bigger is not necessarily better (see ECONOMIES OF SCALE FOR THE INDIVIDUAL FIRM).

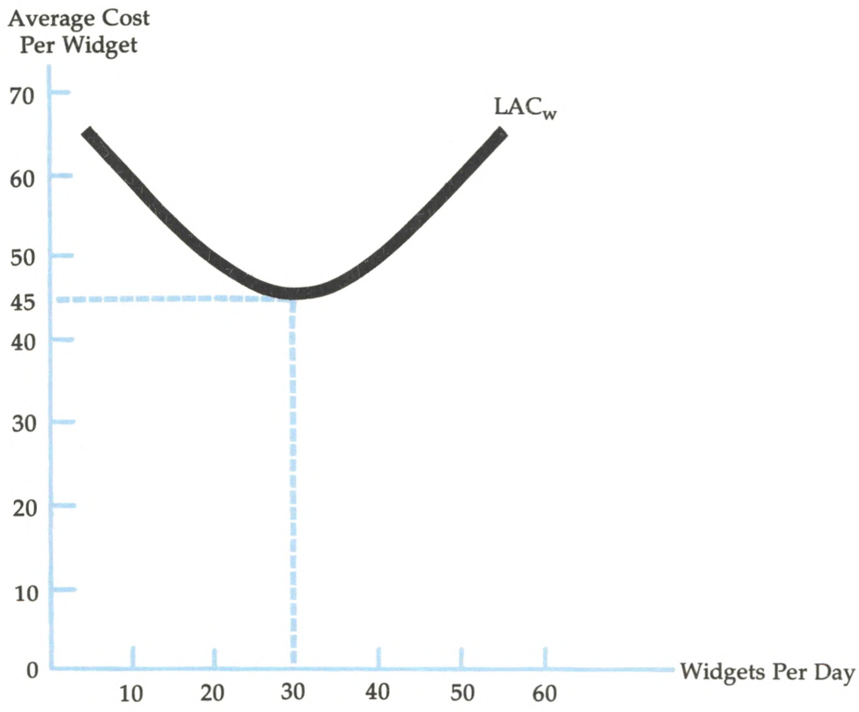
Just as internal economies of scale lead to increased productivity as a firm grows up to a point, *external* economies of scale may also increase a firm's productivity up to some point as well. Economists have identified two such types of external economies of scale, or agglomeration economies. The first type, *localization economies,* depends not upon the size of any one firm in an industry, but upon the size of the firm's industry in a given city. That is, as more and more firms in a given industry locate in a city, each firm's productivity increases. The second type, *urbanization economies,* does not depend upon the size of any one firm in the city, or upon the size of its industry in that city, but upon the overall size of the city itself. That is, as more and more firms of any sort locate in a city, the productivity of each firm increases.[1]

Just as in the case of the growth of an individual firm, growth of an industry in a given city or growth of the city itself increases firms' productivity *only up to some point.* Growth brings not only greater efficiency, but also problems, such as congestion, that may eventually balance or outweigh the efficiency gains from size. When size becomes a hindrance rather than a help, firms in a city experience what is called "diseconomies of scale."

The Size of a Firm's Industry in a City Matters... The origin of an industry in a particular city could be the result of natural resources or

[1]The expressions "city," "urban," "urban areas," "metropolitan area," and their adjectives are being used to designate a metropolitan statistical area (MSA). In general, MSAs are statistical constructs used to represent integrated labor market areas. They typically are geographic areas combining a large population nucleus with adjacent communities that have a high degree of economic integration with the nucleus.

Economies of Scale for the Individual Firm



The notion of economies of scale for a firm is easily illustrated by looking at the long-run average cost curve (LAC_w) for a hypothetical firm, Original Widgets. The long-run average cost curve shows how a firm's cost of production changes as it varies all inputs, including its plant size, or scale of operation. Economies of scale enable the firm to produce large outputs at lower average cost than small outputs. For example, a large financial outlay is usually required to commence production at all. The larger the output, the less is this fixed outlay per unit of product. In theory, a firm's long-run average cost curve is "U-shaped;" as output increases, average cost decreases up to some point and then increases.

To produce 10 widgets a day, Original Widgets' cost per widget is $60. By producing twice as many widgets per day, Original Widgets cuts its average cost to $50 per widget. If Original Widgets again increases production, to 30 widgets per day, its average cost again declines, though not by as much, to $45 per widget.

Up to this point, Original Widgets has enjoyed increasing returns to scale. At 30 widgets per day, the minimum point of its average cost curve, Original Widgets is producing at the point of constant returns to scale. Expansion of widget production beyond 30 units per day results in an increase in Original Widgets' average cost. The average cost of producing 40 widgets per day goes back up to $50, the same as the cost of producing only 20 widgets per day. In other words, Original Widgets is operating with decreasing returns to scale. Decreasing returns may happen when Original Widgets grows so large that it becomes hard to manage effectively. Original Widgets actually does best when it achieves constant returns to scale, or 30 widgets per day at an average cost of $45 per widget.

simply a historical accident. But once an industry develops in a geographic location, individual firms in that industry often reap special benefits by also locating there. Consider for example, California's Silicon Valley, Route 128 near Boston, North Carolina's Research Triangle, and Route 202 in the Philadelphia suburbs, four areas where the computer industry has concentrated. Computer manufacturing firms occasionally require highly specialized workers who maintain and repair computer manufacturing instruments. A computer firm located far from one of the industrial clusters would need to employ full-time computer repair specialists, or else spend considerable time and money bringing them from a distance when they are needed. But when firms cluster together, their combined needs for the repair of their instruments can support at least one firm that specializes in instrument repair. Thus, those services become available at lower cost from a local firm. All the computer firms in the cluster can enjoy a lower average cost of production by contracting for these specialized services only when they are needed.

Of course, computer manufacturing firms that cluster together conceivably share a number of other inputs. For example, these industrial concentrations tend to contain common pools of specialized workers that any one firm in the industry can draw upon when it wants to expand its work force. They also typically contain suppliers of component parts, such as computer chips, and other intermediate inputs that are used by many firms in the industry.

Localization economies undoubtedly played a significant role in the concentration of the motion picture industry in Los Angeles, the auto industry in the Detroit area, and the steel industry in the Pittsburgh region. While a localized input such as ore deposits, or a large body of water, may have been important in getting these industries started, localization economies have been a factor in maintaining these concentrations. In 1985 over half the steel production in the U.S. was concentrated in three states, Pennsylvania, Ohio, and Indiana, and in 1986 Michigan accounted for about 44 percent of total employment in the auto industry.

. . . And So Does the Size of the City Itself. Just as some kinds of business, such as the repair of computer instruments, are found only where specific industries concentrate, other activities, such as financial and business services, are generally found only in urban areas. In some cases, only a large urban setting can provide a sufficient client base for these specialized firms to flourish. Access to these types of specialized services in a city gives rise to the economies of scale that are external to any one firm and to its industry—urbanization economies. Urbanization economies involve the more general cost savings that a firm in any industry may receive by locating in a metropolitan area. For example, urban areas provide wholesaling facilities that reduce the level of inventories any one firm needs on hand. Urban areas also provide access to large and varied labor pools, and to accounting, data processing, legal, and other specialized business services. A *Wall Street Journal* story (July 7, 1987, p. 1) reports an interesting example. A bicycle manufacturer in suburban Boston was "too small to have a full-time chief financial officer, but big enough to have some of the same problems that confront far larger companies." However, it was able to find a local firm that provides financial managers who spend part of each week "doing what CFOs are supposed to do: prepare budgets, project sales, negotiate with banks, and figure out how to cope with the sagging dollar."

The degree of urbanization economies depends upon the number of firms in a city, regardless of what industry they represent. Some of the advantages that a firm gets by locating in one of the nation's largest cities, such as New York City, Los Angeles, or Chicago, could not be realized by locating instead in much smaller cities such as Akron, Ohio, or York, Pennsylvania. New York City not only has many banks, investment houses, advertising agencies, and law firms, but it is large enough to maintain highly specialized varieties of these types of firms. In addition, New York City's labor market is so large that it offers not only a large number of placement firms, but also a large number of agencies that specialize in particular kinds of personnel.

HOW MUCH DIFFERENCE DOES SIZE MAKE?

Economists have measured agglomeration economies by applying the notion of a production function to metropolitan areas. A production function shows the relationship between the inputs of production (labor, land, capital, and so on) and output. The production function for an individual firm will show whether proportionate changes in all its inputs lead to a proportionate increase in output (constant returns), a more than proportionate increase (increasing returns), or a less than proportionate increase (decreasing returns). If increasing returns to scale or agglomeration economies exist in a city, we would expect to find that a proportionate change in all inputs in a city would result in a more than proportionate increase in output (see AGGLOMERATION ECONOMIES LOWER THE AVERAGE COST OF PRODUCTION, p. 8).

Empirical analysis of agglomeration economies has had to deal with two data problems. First, data on the stock of capital at the metropolitan area level are simply not available.[2] Fortunately, a production function technique has been developed that permits the estimation of economies of scale without the need for data on the capital stock.[3] Second, data on industries other than manufacturing are sparse. Therefore, research has had to focus almost exclusively on manufacturing industries in the past 15 years to determine whether agglomeration economies are a fact of economic life for U.S. cities.

Two studies from the 1970s focusing at the industry level take somewhat different approaches to estimate the degree to which agglomeration economies exist for manufacturing in U.S. cities. Daniel Shefer looks at 20 industries in a cross section of cities (ranging from 26 cities in the leather industry to 62 cities in the printing and publishing industry) in the years 1958 and 1963.[4] He finds evidence of economies of scale for urban manufacturing industries in both years. For example, for the primary metal industry in 1963, he estimates that, on average, a 1.0 percent increase in all inputs used by this industry in a city would result in a 1.12 percent increase in output. One limitation of the Shefer study is that we do not know to what extent his estimates reflect internal or external economies of scale.

In a more recent study, Gerald Carlino extends the analysis of agglomeration economies. He estimates economies of scale for 19 industries in each of 68 metropolitan areas over the period 1957-1972.[5] He derives a single measure of overall returns to scale in each industry in each city over that period. He then analyzes these industry-specific measures across cities to determine the extent to which overall economies of scale are related to internal economies of scale, localization economies, and urbanization econo-

[2]Some researchers have put together estimates of capital stocks. However, their results are not strictly comparable to those reported here, though the general direction of the results is the same. See David Segal, "Are There Returns to Scale in City Size?" *Review of Economics and Statistics* 58 (1976) pp. 339-350. Segal analyses the change in urban productivity with city size but does not focus on agglomeration economies, which is one component of city productivity. He finds that on average cities with over 2 million people are 8 percent more productive than cities with under 2 million people. See also Patricia Beeson, "Total Factor Productivity Growth and Agglomeration Economies in Manufacturing, 1959-73," *The Journal of Regional Science* 27 (1987) pp. 183-190. Since Beeson uses *state* level data her findings are hard to compare with those reported here. Beeson uses a capital stock series developed in Lynne Brown, Peter Mieszkowski, and Richard Syron, "Regional Investment Patterns," *New England Economic Review*, (July/August 1980) pp. 5-23.

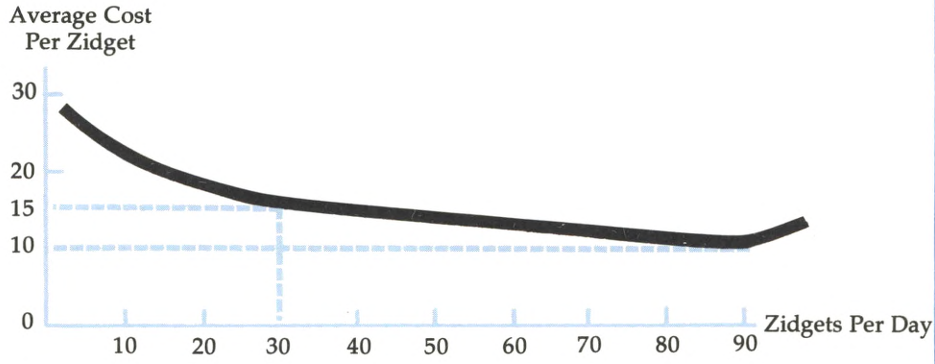
[3]The technique involves estimating a wage equation. It is assumed that, since workers are paid according to their productivity (that is, there is perfect competition in local labor markets), wages and the demand for labor reflect the advantages of agglomeration economies.

[4]Daniel Shefer, "Localization Economies in SMSAs: A Production Function Approach," *Journal of Regional Science* 13 (1973) pp. 55-64.

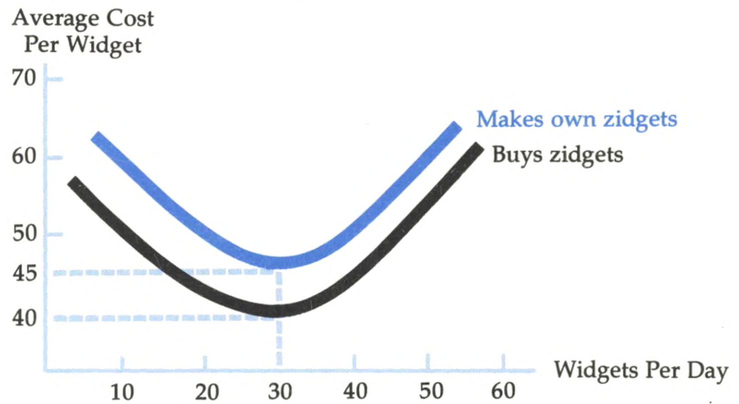
[5]Gerald A. Carlino, "Increasing Returns to Scale in Metropolitan Manufacturing," *Journal of Regional Science* 19 (1979) pp. 343-351.

Agglomeration Economies Lower the Average Cost of Production

Letting the Specialists Produce the Zidgets . . .



. . . Lowers the Overall Cost of Producing Widgets



To see how agglomeration economies can lower a firm's long-run average cost, we can return to Original Widgets and assume that a crucial part of making a widget involves a fitting called a zidget (in reality, this crucial factor may be repair services, accounting, financial and legal services, computer programmers, and so forth). The top panel shows the long-run average cost of making zidgets, which involves substantial economies of scale. At 30 zidgets per day the cost is $15 per zidget, but when 90 zidgets are produced, the cost per zidget drops to $10.

Original Widgets cannot take full advantage of the economies of scale of zidget-making *internally*, because it only needs 30 per day for its widget production. But, suppose the local widget industry expands, say to three widget firms producing a total of 90 widgets per day. Now a separate firm, Acme Zidgets, can take advantage of the economies of scale of zidget production and supply them to all three widget manufacturers in the area. The result is shown in the second panel, where Original Widgets (as well as the other two widget firms) enjoys the cost-savings due to agglomeration economies because its average cost of producing 30 widgets per day could drop by as much as $5 (from $45 to $40).

mies by industry. The results strongly suggest the importance of external economies of scale for urban manufacturing firms. He finds that urbanization economies are the more general source of external economies, since they are indicated for 12 of the 19 industries studied.[6] For five of the remaining industries, localization economies are an important source of external economies of scale. Of the two remaining industries, internal scale economies are indicated in one case, and no significant source of economies of scale were found in one industry (see SOURCES OF ECONOMIES OF SCALE IN SELECTED INDUSTRIES, p. 10).

Using the same techniques with which they measured agglomeration economies at the industry level, Shefer and Carlino also estimate the degree to which agglomeration economies exist for manufacturing in general in U.S. cities. This provides another measure of urbanization economies. In the same study covering 67 cities in 1963, Shefer finds that a 1.0 percent increase in inputs used in urban manufacturing results in about a 1.2 percent increase in urban manufacturing output, on average. Carlino obtains an estimate of urbanization economies in manufacturing city by city. His study includes 82 metropolitan places during the period 1957 to 1977.[7] He finds that agglomeration economies also tend to increase with city size, up to some point. For example, a 1.0 percent increase in all inputs resulted in a 1.9 percent increase in output in Peoria, a 1.4 percent increase in output in Cincinnati, a 1.3 percent increase in output in both Kansas City and St. Louis, and a 1.2 percent increase in Boston's output. But estimates for Philadelphia, the fourth largest metropolitan area in the U.S. in 1980, indicate only about a 1.0 percent increase in output (that is, constant returns).

Why would a large metropolitan area such as Philadelphia, which contained approximately 4.7 million people in 1980, offer constant returns to its manufacturing firms on average while, say, Peoria, which contained less than 400,000 people in 1980, offers substantial returns to scale on average? The answer lies in the costs to both firms and households that result from increased urban size.

FACING THE COSTS OF GROWTH

The positive effects of agglomeration economies make up one side of the urban size ledger. The negative effects of congestion on households and firms brought on by city growth (agglomeration diseconomies) make up the other. The growth of cities is influenced by both forces, with agglomeration economies encouraging growth, and agglomeration diseconomies discouraging it.

Higher Transportation Costs . . . After some point, further increases in the number of people and firms residing in a metropolitan area tend to clog its roads and transportation network and increase the average time and cost of transporting goods and commuting either to work or to leisure activities. In addition, as a metropolitan area grows, its boundaries may spread out, which increases both the time and distance of the average commute. As a result, households will have to spend more annually for gasoline and auto maintenance, and they may even need to purchase a second car.

. . . Higher Rents . . . Most commuting to work involves trips to and from a metropolitan area's

[6]Two other studies using a somewhat different technique from the one discussed here have found that localization economies are dominant in urban manufacturing. See, J. Vernon Henderson, "Efficiency of Resource Usage and City Size," *Journal of Urban Economics* 19 (1986) pp. 47-70, and Ronald Moomaw, "Agglomeration Economies : Urbanization or Localization?" unpublished manuscript (1987). These findings may follow from the fact that they use only industry level data.

[7]Gerald Carlino, "Manufacturing Agglomeration Economies as Returns to Scale," *Papers of the Regional Science Association* 50 (1982) pp. 95-108. A number of studies since the mid-1970s have shown that productivity in general increases with the size of a city at least over the observed range of city sizes. These studies have found that a 3 to 6 percent increase in city productivity is associated with every doubling of city size. For a survey of this literature, see Ronald Moomaw, "Spatial Productivity Variations in Manufacturing: A Critical Survey of Cross-Sectional Analyses," *International Regional Science Review* 8 (1983) pp. 1-22.

Sources of Economies of Scale in Selected Industries

These 19 industrial groupings represent the two-digit Standard Industrial Classifications as defined by the U.S. Office of Management and Budget. Each consists of a fairly broad aggregate of establishments, each of which may derive different benefits from localization economies and urbanization economies. For example, the classification Electrical Machinery includes establishments engaged in manufacturing equipment for the generation, storage, transmission, and transformation of electrical power, establishments manufacturing computers and related products, and firms manufacturing household appliances. Thus, the finding that a particular two-digit industry does or does not depend upon a particular kind of agglomeration economy may not apply to all of its component establishments. The data used in this study are averaged over the period 1957-1972.

Industry	Internal	External	
		Localization	Urbanization
Food Products	Yes	No	Yes
Textiles	No	No	Yes
Apparel	No	No	Yes
Wood	No	No	Yes
Furniture	No	No	Yes
Paper	No	Yes	Yes
Printing and Publishing	No	No	Yes
Chemical	No	No	No
Petroleum and Coal	No	No	Yes
Rubber and Plastics	Yes	No	Yes
Leather	No	No	Yes
Stone, Clay and Glass	No	Yes	No
Primary Metal	Yes	No	Yes
Fabricated Metal	No	No	Yes
Nonelectrical Machinery	No	Yes	No
Electrical Machinery	No	Yes	No
Transportation Equipment	No	No	No
Instruments	No	Yes	No
Others	No	No	Yes

SOURCE: Compiled from G.A. Carlino, "Increasing Returns to Scale in Metropolitan Manufacturing", *Journal of Regional Science* 19 (1979) Table 2.

downtown, or its central business district. Many firms seek these central locations, in part, because they offer agglomeration economies. This competition will increase business rents. As households attempt to locate near these large centers of economic activity to avoid long commutes, they bid up residential rents as well. As a result, rents in a metropolitan area tend to reflect the proximity of a parcel of land to its central business district. Moreover, rents in an entire metropolitan area tend to be driven up by the growth of households and firms in that area.

. . . And Higher Wages. Workers in large cities will demand higher wages in order to offset these increases in transportation costs and rents.[8] As a result, wages tend to increase with metropolitan size. Firms are able to pay these higher wages to workers to the extent that agglomeration economies have made workers

[8]For a fuller discussion that includes analysis of the effects of local amenities and disamenities on wages, see Sherwin Rosen, "Wage-Based Indexes of Urban Quality of Life," in *Current Issues in Urban Economics*, Peter Mieszkowski and Mahlon Straszheim, eds., (Baltimore: The Johns Hopkins University Press, 1979).

more productive. But there is a limit to a firm's ability to compensate its workers for these higher living costs.

Finding the Balance: Is There An Optimal City Size? The agglomeration diseconomies reflected in higher transportation costs, higher rents, and higher wages serve to increase the unit cost of production for firms. As long as these additional costs are offset by increased productivity, firms will be willing to pay them, and a city will continue to grow. When the unit cost-saving from the agglomeration of people and firms is just offset by the increased cost due to agglomeration diseconomies, a city has reached what economists call its optimal size. At the optimal size, the average cost of production is minimized.

While the notion of an optimal city size has been addressed in a number of studies, it has proven hard to identify precisely for any city. Part of the reason is that a city's optimal size will depend on its mix of industries, its proximity to other cities, its rate of technical change and the level of its infrastructure. Since the cost of labor and land and the advantages of agglomeration economies vary with city size, firms' decision about locating in cities of particular sizes will depend on how much they use labor and land, and how much they would gain by taking advantage of agglomeration economies. The estimation of the optimal size for individual cities has not been attempted because the size of the population in most cities has not varied substantially during the period for which data are available.[9]

HOW CAN POLICYMAKERS ENHANCE PRODUCTIVITY?

Individual firms that have incentives to exploit agglomeration economies are guided by the "invisible hand" of the marketplace to locate near other firms in the same industry or in areas where there is a general concentration of economic activity. Local policymakers have a major role in "lending a hand" to accommodate agglomeration economies by providing public infrastructure. In this sense public infrastructure and private capital are complementary inputs to local production. Local policymakers have an additional role to play in enhancing the productivity of firms by investing in the education of the city's workers.[10] This sort of investment in education can result in what economists call increased "human capital."

Investment In Public Infrastructure. When a city is growing rapidly because it offers net agglomeration economies to firms in a number of industries, local planners need to make sure that the city's public infrastructure keeps in step with private growth. If local infrastructure is not growing fast enough, the area could become congested more rapidly, leading to a more rapid increase in wages, rents, and transportation costs. Such a situation could halt the growth of an area. After some point, additional public infrastructure is necessary for future growth to occur.

In a recent study, Randall Eberts measured the level of public infrastructure for 38 metropolitan places for the time period 1958-1981.[11]

[9]Economists have, however, estimated an optimal size for an *average* city based on economies of scale in manufacturing using cross-sectional data. See, for example, Gerald A. Carlino, "Manufacturing Agglomeration Economies as Returns to Scale," *Papers of the Regional Science Association* 50 (1982) pp. 95-108 who finds the optimal size to be around 3.5 million people.

[10]Factors other than those discussed here could affect city productivity. They include the characteristics of a city's work force other than educational and skill attainment, local policies and regulations, research and development spending, unionization rates, and environmental considerations. While these factors may determine differences in city productivity, little, if any, research has been conducted on these issues.

[11]Randall Eberts, "Estimating the Contribution of Urban Public Infrastructure to Regional Growth," Working Paper 8610, Federal Reserve Bank of Cleveland (1986). Eberts estimates the level of public infrastructure by summing up the past investments made to the stock of infrastructure in each of these metropolitan places, after adjusting these stocks for depreciation and discard. He uses a pooled cross-section of time-series approach to derive an average estimate of the effect of infrastructure on productivity across these 38 metropolitan places.

As with the studies that examine the effect of agglomeration economies on productivity, Eberts considers public infrastructure an input, together with labor and private capital, in a citywide manufacturing production function. While his method of estimation gives no particulars about specific cities, Eberts finds that a doubling of public infrastructure would lead to a 4 percent increase in manufacturing output on average in his sample of 38 metropolitan places.

Investment In Education. Besides determining the quality of local infrastructure, local policymakers in the U.S. have a great deal of influence on the skill level of the work force because they control the public education system. These investments in human capital lead to increased city productivity not only because education makes a city's work force more employable but also because education introduces a city's workers to new techniques and skills. For example, many high schools throughout the country have developed programs in computer literacy.

John Mullen and Martin Williams consider these issues in manufacturing for a sample of 29 metropolitan places during the 1958-1978 time period.[12] They compute the portion of a city's growth of manufacturing output that can be accounted for by that city's increases in the number of workers and capital that took place during the period. The growth of a city's output beyond that explained by the increases in capital and labor they attribute to technical progress. They then decompose this measure of technical progress into that which is due to better workers (embodied in labor) and that which is due to better capital (embodied in capital). They find that across metropolitan places, technical progress embodied in the labor force was a more important source of productivity growth than technical progress embodied in private capital. This study suggests that local policies that increase the educational attainment and skill levels of its work force are highly worth pursuing.

Where Best to Put the Effort? Much has been written about the decline of large American cities — the urban blight, the crime, the many negative but very tangible and visible features of American urban life. But large cities have existed and will continue to exist at least in part because they tend to make workers and other factors of production more productive, as various studies have shown. Local planners need to recognize the fact that city size matters, for if they allow infrastructure and schools either to remain as they are or to decay, they will fail to exploit to the fullest the growth that agglomeration economies provide.

[12]John Mullen and Martin Williams, "Technical Progress in Urban Manufacturing," *Journal of Urban Economics* (forthcoming). One problem with this approach to measuring technical progress is that it fails to account for the growth in output that is due to agglomeration economies. As a result, some of the increase in productivity that is attributed to technical progress may be due to agglomeration economies.

Commuter Rail Ridership: The Long and the Short Haul

*Richard Voith**

INTRODUCTION

Many American cities have commuter rail systems which, in addition to serving their riders, are intended to benefit the region as a whole by reducing congestion and air pollution, enhancing economic development, and providing transportation services to the poor. The degree to which these potential benefits are realized depends upon the number of riders the system can attract. A commuter rail system with little patronage cannot contribute much to congestion relief or air pollution abatement.

Demand for commuter rail transportation, like the demand for any service, depends upon its price, the price of alternatives, and the quality of the service. Unlike most other services, however, prices or fares in the regional public transportation industry are determined not in the marketplace but by a public authority. Most public transportation systems, including commuter rail systems, depend on state and local governments for subsidies, as fares cover only a portion of the operating cost. Fares, the quality of service, and ultimately the level of ridership,

*Richard Voith is an Economist in the Urban and Regional Section of the Research Department of the Federal Reserve Bank of Philadelphia.

will depend on the level of subsidy available for public transportation and on how that subsidy is allocated throughout the service area. When state and local governments decide how much subsidy to provide, they walk a fine line between allocating enough funds to reap the benefits of public transportation and keeping enough budgetary pressure on the transit authority to provide the service in a cost-effective manner.

If lower subsidies induce the transit authority to produce transportation services more efficiently, that is clearly beneficial. Transit authorities, however, often respond to budgetary shortfalls by increasing fares, reducing service, or both. While such actions balance the budget in the current year, they can lead to problems in the future. In the short run, increases in price and reductions in service have a relatively small impact on ridership. But, in the long run, consumers can exercise more options among their commuting alternatives; therefore, ridership may decline after the initial impact of the price increase or service reduction, leaving the system with lower and lower farebox revenues. The difference in commuters' short-run and long-run responses to changes in price and service levels may help explain the familiar cycle of service reductions, increasing fares, and falling ridership often observed in the public transportation industry.

THE EVOLUTION OF COMMUTER RAIL DEMAND

Consumers are the ultimate judges of public transportation policies, and they evaluate public transportation relative to the price and quality of other alternatives. Important elements in the quality of commuter rail transportation are frequency of service, speed, reliability, and factors affecting comfort, such as crowding and cleanliness. Changes in the price or characteristics of the rail system (or of competing means of transportation) will affect the choices of some consumers immediately, while others will be affected only after some lag as they make long-term decisions.

In the short run, a consumer faces a fairly narrow set of alternative types of transportation and will choose the most attractive among them to get from place to place. For example, he can choose to drive if he owns a car, or take the train or a bus from home to his place of work. In the short run, the consumer's transportation alternatives themselves and the origins and destinations of trips cannot readily be changed.

Over the longer term, however, a consumer can change his transportation alternatives by making investments, such as purchasing a car or perhaps a second car. He might be able to join a car or van pool to reduce the cost of private transportation. He can even change the origin and destination of his commuting trips by moving or changing employment. Transportation is often a major consideration in such a change. Thus, in the long run a consumer has considerably more options in responding to changes in the relative prices and qualities of various transportation alternatives.

It is not just current price and quality that affect these long-run decisions but future considerations as well. If there is a great deal of uncertainty about the price or the existence of the commuter rail service in the future, the potential benefits of that system are discounted in the consumer's long-term decision.

Taken together, the short-run and long-run decisions of consumers in the entire region determine the evolution of ridership over time. If the price and quality of train service make it an attractive alternative, people and firms are likely to make long-term location and investment decisions that will lead to high levels of ridership in the future. Areas well served by the system will grow and develop. Individuals who work along the train lines will sort themselves into residential locations that have train service. For instance, most systems have a hub in the center of the city so that people who work there will be more likely to live in areas with train service, which will probably raise property values there. On the other hand, people who work where train service is not available will choose to live in areas that are not near train stations to avoid paying higher housing prices. Locations well served by the train, therefore, will have a disproportionately large number of people who routinely travel by train. Hence, ridership will be high.

On the other hand, if the quality of service is poor, or too expensive, or if future subsidies are uncertain, individuals and firms will not weigh the possibility of future train service heavily in their investment and location decisions. People will invest more in automobiles, making it less likely that they would choose to ride the train in the near future, even if the price and quality of the train service were improved. People may choose to live in areas not served by the train even if their job location has train service. Employers may choose locations not served by the train. As a result, where people live and work would not be consistent with high future ridership.

Setting the Public Transportation Budget. Since the transit authority and state and local governments together choose the prices and service levels, they influence the evolution of demand. The transit authority actually sets the fares and service levels within the framework of a balanced budget. Its operating expenses cannot exceed its revenues, which include both proceeds from the farebox and government subsidies. But the authority can only go so far in balancing the budget by cutting operating expenses or increasing fares. Operating expenses cannot be reduced if they result in service levels that are inadequate to sustain consumer demand. And the amount of revenue available from the farebox is limited because riders can opt for other means of transportation if fares are too high.

The other source of revenue for balancing the budget, namely the amount of subsidy available, is a matter of public policy. When state and local governments choose the level of subsidy for public transportation, they weigh a myriad of economic and political considerations.[1] In addition, they often use legislative review of the subsidy and budgetary restraint to induce the transit authority to minimize waste.[2] The share of expenditures covered by the farebox, or operating ratio, is a common measure of the performance of public transportation authorities. Achieving a high operating ratio, however, may not necessarily coincide with achieving the lowest subsidy cost per passenger. It is possible, for example, for a transit authority to attain a very large share of revenue from the farebox by charging high fares, while having relatively low ridership. In this case the subsidy would benefit few riders, and the benefits in terms of traffic congestion relief would be small. Forcing high fares through low subsidies may result in high subsidy costs on a per rider basis. Since the public benefits of the system depend on the level of ridership, a better goal for policymakers may be to choose the subsidy that minimizes subsidy cost per rider.

The "Catch-22" of Public Transportation. There is a trade-off between the reduced waste induced by budgetary restraint and the adverse long-run impacts of higher prices and lower service which may result from a low level of subsidy. In the short run, ridership may not change much in response to changes in price and service levels. Thus, service cuts and fare increases may balance the budget in the current period. But, because of the effect of fares and service levels on people's long-run decisions, the loss in ridership and corresponding decline in farebox revenue resulting from changes in prices and service may be much greater in the long run. In economic terminology, demand is more elastic in the long run than in the short run. (See SHORT-RUN AND LONG-RUN ELASTICITIES, p. 16.)

[1]For a discussion of the role of public investment and productivity, see Gerald A. Carlino, "Productivity in Cities: Does City Size Matter?" this issue of the *Business Review.*

[2]Several studies have noted a correlation between higher levels of government transit subsidies and higher transit worker wages and lower productivity. See J. Gomez-Ibanez, "The Federal Role in Urban Transportation," in *American Domestic Priorities: An Economic Appraisal,* John M. Quigley and Daniel L. Rubinfeld, eds. (Berkeley: University of California Press, 1985) pp. 183-223.

Short-Run and Long-Run Elasticities

Economists often express the change in demand for a product in response to a change in its price or some other factor in terms of "elasticities"—the percentage change in one thing divided by the percentage change in another. Consider, for example, the price elasticity of train ridership:

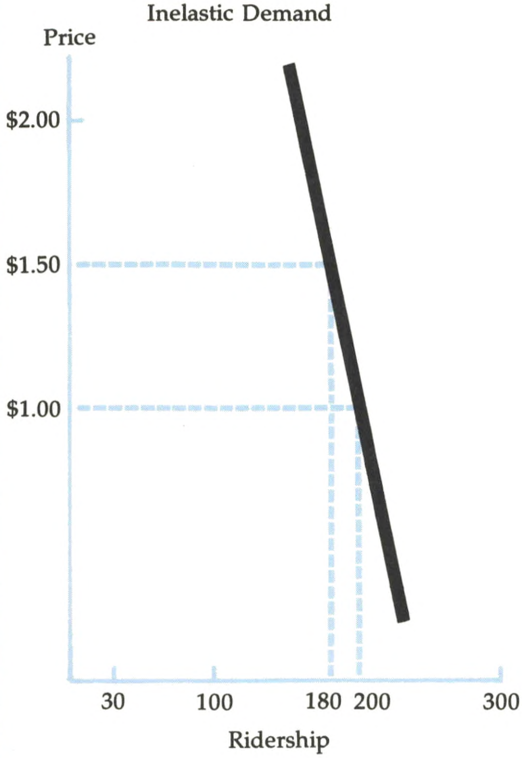
$$\text{Elasticity} = \frac{\text{Percent Change in Ridership}}{\text{Percent Change in Price}}$$

If this ratio is more than 1 (ridership, in percentage terms, changes more than price in percentage terms), then price demand is "elastic." If the ratio is less than 1 (ridership, in percentage terms, changes less than price in percentage terms), then price demand is said to be "inelastic."

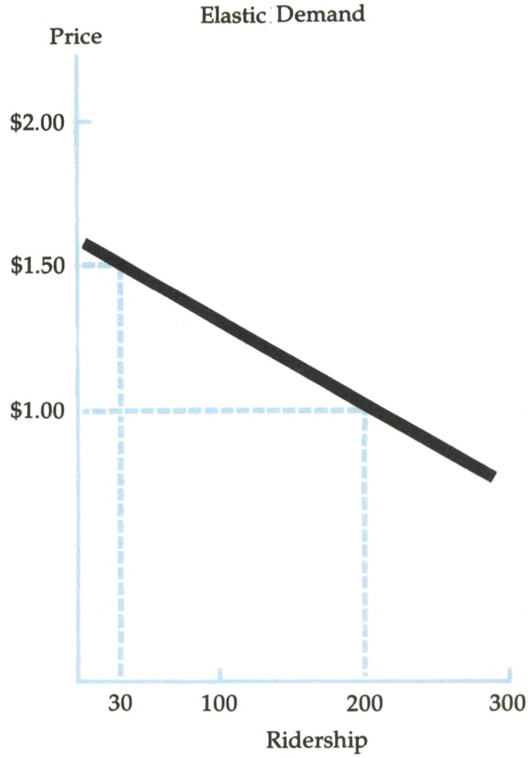
It is a general economic proposition that demand is more elastic in the long run than in the short run. Thus, in some cases, price increases may produce more revenue in the short run, but in the long run during which people have more time to exercise other options, price increases may lead to declines in total revenue.

The graphs below compare examples of elastic and inelastic demand curves to illustrate their effects on total revenues. In both cases, when the railway fare is, say, $1.00 per ticket, the quantity demanded is 200, and total revenue is $200. But, if the fare goes up to, say, $1.50 per ticket, the effect on total revenue is very different depending on the elasticity of demand. Where demand is inelastic, total revenues increase from $200 to $270 even though ridership declines somewhat, from 200 to 180. But where demand is elastic, ridership falls so much—from 200 to 30—that revenues are only a small fraction of what they originally were, falling from $200 to $45.

Inelastic Demand



Elastic Demand



If long-run demand is significantly more elastic than short-run demand, then price hikes and service cuts are likely to result in higher long-term subsidy costs per passenger. Because high fares or poor quality service lead people to look for alternatives to the train, revenue from the farebox falls but the large fixed costs of the rail system remain unchanged; thus the governments' cost per rider increases. Since transportation policies are an important factor shaping the development of a region, the long-run and short-run effects of changes in fares and quality of service should be important considerations of both the transit authority and its state and local subsidizers. A prerequisite for formulating rational transportation policy is knowledge of the short-run and long-run impacts of price and service changes.

ANALYZING THE DEMAND FOR COMMUTER RAIL TRANSPORTATION

From a planning perspective, transit authorities need to know how much the demand for rail transportation is affected by changes in price, quantity, and quality of service. Measuring the total effects of these changes is difficult because the level of ridership depends not only on the price and attributes of train service but also on a number of other factors, such as the number of potential customers, their transportation preferences, their investments in private transportation, and the price and quality of alternatives to the train, such as buses and van pools. The size and makeup of the potential pool of riders play important roles in the level of ridership at any particular location.

Most studies have focused only on the short-run impacts of price and service characteristics on demand, assuming that the choices commuters have now are the only ones available. By observing the choices of many individuals, each facing different circumstances in terms of the prices and attributes of the alternative modes of transportation, the short-run impact of changes in prices and service on their transportation choices can be measured.[3]

Since these short-run analyses do not take into account the transportation system's impact on individuals' long-term choices, and hence its impact on the potential pool of riders, they underestimate the total impacts of price and service changes. To predict the total impact of changes in the price and service levels of the train system on ridership, one must take into account the effects which may not occur instantaneously, but rather gradually as such changes affect the locational distribution of the regional population and the investment decisions of that population.[4]

By examining the evolution of ridership at particular locations in a region over a period of several years, one can estimate both the long-run and short-run impacts of price and service changes.[5] Fortunately, data are available for this type of analysis from the Southeastern Pennsylvania Transportation Authority (SEPTA) rail system in the Philadelphia area (see the Appendix, pp. 22-23, for technical details of the study). From 1978 to 1986 changes in ridership, prices, and service have varied considerably from station to station in the SEPTA system (see TRENDS IN SEPTA COMMUTER RAIL RIDERSHIP, pp. 18-19).

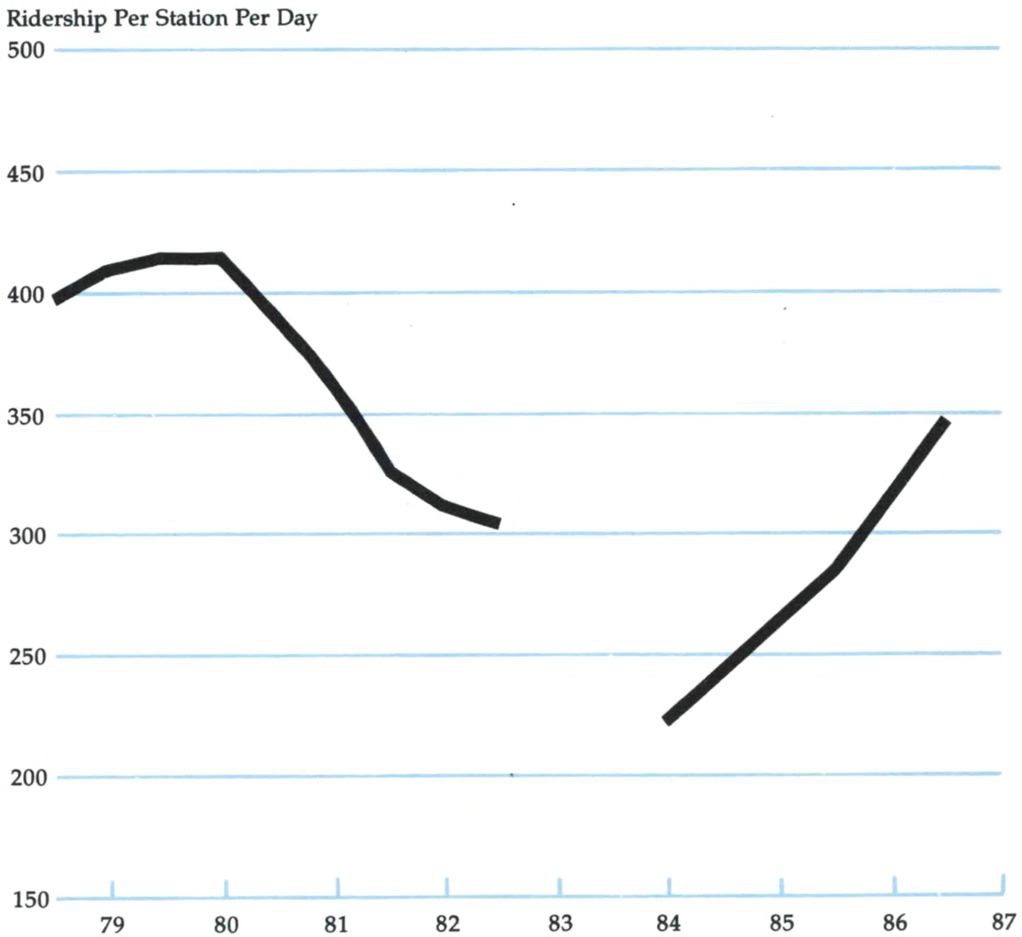
[3]The theory and methods of empirical analysis of individual choice of travel mode in the short run are based on the pioneering work of Daniel McFadden, "Conditional Logit Analysis of Quantal Choice Behavior," in *Frontiers of Econometrics*, Paul Zarembka, ed. (NY: Academic Press, 1974) pp. 105-142.

[4]Mateen Thobani, "A Nested Logit Model of Travel Mode to Work and Auto Ownership," *Journal of Urban Economics*, 15 (1984) pp. 287-301, analyzes the joint decision of purchasing a car and choice of travel mode as functions of the price and attributes of the public transit system. Alex Anas, "Estimation of Multinomial Logit Models of Joint Location and Travel Modal Choice from Aggregated Data," *Journal of Regional Science*, 21 (1981) pp. 223-242, examines the transit system's impact on residential location.

[5]A complete discussion of this methodology is discussed in Richard Voith, "Determinants of Commuter Rail Ridership: The Long and Short Haul," Federal Reserve Bank of Philadelphia Working Paper (forthcoming). A more complete discussion of the data is contained there as well.

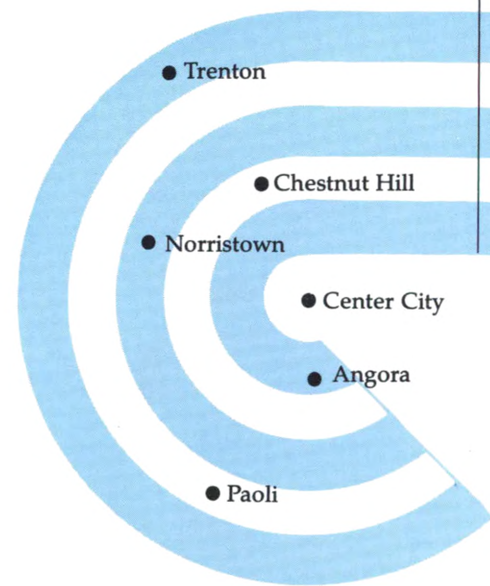
Trends in SEPTA Commuter Rail Ridership 1978-1986

The Overall Trend



From 1978 to 1980, ridership rose slightly to its peak, but in the next two years declined rapidly as fares increased and service levels fell. In 1983, SEPTA took over operation of the system from Conrail and, in an effort to reduce costs, endured a strike that lasted over three months. The gap in the data is a result of the strike; fall 1982 was the last pre-strike observation and spring 1984 was the first post-strike observation. By spring 1984, ridership had fallen dramatically to its all time low. Since 1984, ridership has rebounded to about 80 percent of its 1980 peak.

The Zone-By-Zone Trend



Percent Change, 1978-1986

Zone	Ridership	Price	Number of Trains
5	5.7	97.3	25.7
4	- 6.0	126.0	12.1
3	- 6.5	128.0	2.9
2	-32.2	176.5	-19.7
1	-41.0	151.7	-14.6

The aggregate figures mask significant differences over time when the data are broken down by fare zone. Zones 1 and 2, which are the closest to the center of the city, have had dramatic declines in ridership, while ridership fell slightly in zones 3 and 4 and increased in zone 5. The dramatic fall in ridership in the interior zones was accompanied by significant reductions in the total number of trains and large price increases. On the other hand, in zone 5 where ridership increased, the total number of trains increased by 25 percent and the price increase was much smaller.

While some of the ridership loss in the close-in zones, especially zone 1, may have been caused by population declines in the city of Philadelphia (unrelated to the changes in the transportation system), these declines are very small relative to the magnitude of the decrease in ridership. It appears that much of the ridership loss is a result of the price increases and service reductions. One might expect these areas to be especially sensitive to price, because alternative forms of public transportation—buses and subways—are available. Also there has been significant improvement in the quality of the bus and subway system over this period. In the more distant zones, population growth should have provided natural growth in ridership for the commuter rail system. However, with employment booming in these outlying areas, many people now both live and work there. The increase in ridership in the most distant zone indicates that the increase in service and more modest price increases had a positive effect on ridership.

Short-Run and Long-Run Demand Elasticity on the SEPTA Rail System. Econometric analysis of the SEPTA data reveals that the long-run responses of ridership to changes in prices and service attributes are considerably larger than the short-run responses (see Table 1). Short-run responses—those which occur at the time of the change—all proved to be inelastic; that is, the percentage change in ridership is less than the percentage change in price, number of peak or off-peak trains, or speed of the train. As predicted, the estimated total impacts of changes in prices and service attributes are much larger than the short-run impacts—more than twice as large. The analysis further suggests that about half of the total impact occurs within the first year.

In the case of price, the short-run elasticity is about -0.68, meaning that a 10 percent increase in price generates a 6.8 percent decrease in ridership. This estimate is similar to other measurements of the short-term price elasticity of other commuter systems.[6] The long-run elasticity is almost three times as great, at -1.84. To illustrate how these price elasticities could affect revenues (holding everything else constant), suppose SEPTA, which has about 100,000 riders, increased the average one-way ticket price by $0.25 or 9.2 percent (Figure 1). Daily revenue would increase immediately by $8,000 per day. So, because ridership is inelastic in the short run, the transit authority could increase revenues in the short run by increasing fares. But, in the long run, the revenue picture deteriorates. After the first half-year, the increase drops to zero; by the end of the first full year, daily revenue is *reduced* by almost $6,500, and after four years, revenues are below the original levels by over $19,000. Since these elasticities work in the opposite way when the fare drops by $0.25, SEPTA might be able to generate more revenue by lowering prices, provided it can handle the extra passengers.

The budgetary implications of elastic versus inelastic demand are less conclusive in the case of service attributes, since the financial effects of changing the service attributes depend not only on the change in ridership but also on the costs of changing the quality of service. The short-run elasticities for the number of peak trains and the number of off-peak trains are 0.19 and 0.54, respectively, while the short-run speed elasticity is about .24. Since the average number of peak trains in 1986 was 7.6, this implies that an addition of one peak train (a 13 percent increase) will increase peak ridership along that line by 2.6 percent. An additional off-peak train (a 5.3 percent increase) would increase off-peak rider-

[6]See Clifford Winston, "Conceptual Developments in the Economics of Transportation: An Interpretive Survey," *Journal of Economic Literature*, 23 (1985) pp. 57-94, for estimates of the price elasticity of ridership on the BART commuter rail system in San Francisco.

TABLE 1

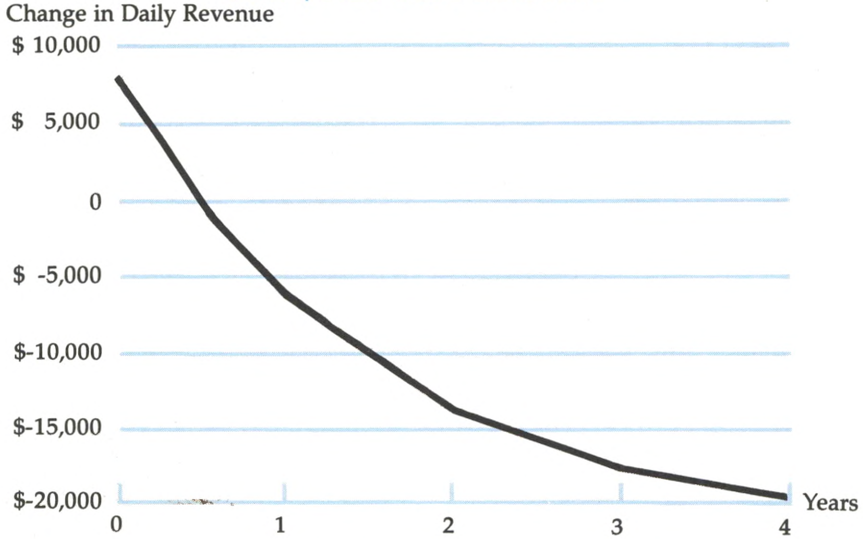
Ridership Is More Elastic in the Long Run

	Short Run	Long Run
Average Nominal Price	-0.68	-1.84
Peak Number of Trains	0.19	0.52
Off-peak Number of Trains	0.54	1.47
Speed	0.24	0.66

NOTE: Elasticities are evaluated at 1986 levels and are derived from the coefficients in the Appendix. For example, the short-run price elasticity is: $\varepsilon_0 = (-41.8 \times 5.4)/333 = -0.68$. The long-run price elasticity is: $\varepsilon_\infty = (1/(1-0.63)) \times -0.68 = -1.84$.

FIGURE 1

The Long-Run Impact of a $0.25 Fare Increase



NOTE: The future revenue impact can be calculated as follows: (1) base year farebox revenue is $F_o = p_oR_o$, where p is price and R is ridership and the subscript denotes the initial period; (2) ridership t periods after the change in price, Δp, is given by $R_t = [1+\varepsilon_t(\Delta p/(p+\Delta p))]R_o$ where ε_t is the t-period price elasticity; (3) the t-period price elasticity can be calculated using the formula, $\varepsilon_t = (\varepsilon_o/(1-\lambda))(1-\lambda^{t+1})$, where ε_o is the short-run elasticity and λ is the lag parameter (0.63) from the Appendix; (4) the farebox revenue in period t is $F_t = R_t(p+\Delta p)$ and the revenue impact is F_t-F_o. Since each period is about eight months long for the available data, the t for a one-year impact is 1.5.

ship by 2.8 percent. Likewise, since the average speed of the system in 1986 was 22 miles per hour, increasing average speed by 10 miles per hour (a 45 percent increase) would result in a 10.8 percent increase in ridership. While not as striking as the long-run price elasticity, the long-run implications of service changes are significant as well. The addition of one peak train would increase peak ridership by 7 percent; increasing speed by 10 miles per hour would increase ridership by 29 percent. An increase in speed would tend to have even greater impacts since the authority could operate more trains with no additional equipment or crews. If the greater speed allowed the frequency of service to go up 30 percent, the combined effect on ridership would be an increase greater than 40 percent.

These estimates indicate that there is considerable scope for SEPTA to increase patronage by increasing speed and frequency and lowering price—if the short-run budget constraint could be loosened and if appropriate investments are made by the transit authority to improve service along the dimensions consumers value. Furthermore, price increases and service reductions may be counterproductive in the long run, even if they do balance the budget in the short run. These actions actually may result in a higher subsidy cost per rider or per passenger mile, though the total subsidy may be lower. This is true not only of the SEPTA system, but of any rail system in which price and service changes have relatively small effects on ridership in the short run and relatively large effects in the long run.

THREE IMPLICATIONS FOR TRANSIT POLICYMAKERS

Three basic policy implications emerge from long-run price and service elasticities that are greater than short-run elasticities. First, transit authorities should closely examine their pricing and service policies to ensure that they are consistent with long-term cost-effective service. This means that the transit authority should actively pursue strategies that encourage development and location decisions that will lead to future ridership.

Second, those who subsidize public transportation should recognize that price increases, service reductions, and uncertainty about the level of future service may have counterproductive effects in the long run. In order to obtain reasonable costs per rider, the subsidy level will have to be large enough to provide service that will induce people to make location and investment decisions that are consistent with public transportation usage. If people are uncertain about the future levels of service, they will insure themselves by becoming less dependent on public transportation, which will lead to lower future ridership.

Because ridership is more responsive to price and service changes in the long run than in the short run, balancing a transit authority's budget through price increases and service reductions may result in future financial difficulties. The findings based on the analysis of data on one commuter rail system (SEPTA) suggest that the long-term impacts may be sufficiently large that further price increases and reduction in the frequency of trains will not improve a transit authority's long-term financial performance.

Finally, because the consequences of price and service changes are not completely manifest in a single budget year, state and local governments should consider alternatives such as multiyear appropriations. In that case, transit authorities could balance their budgets over a longer period rather than in each budget year. A longer planning horizon would allow transit authorities to avoid making short-run decisions which, in the long run, can lower ridership and increase subsidy costs per rider.

Appendix

The estimates of elasticities for the SEPTA system reported in this article are derived from a dynamic fixed-effects model. The basic model of ridership from any location consists of two equations:

$$R^{jt} = h(p_o^{jt}, p_a^{jt}, A_o^{jt}, A_a^{j}; D^{jt})$$

$$D^{jt} = f(p_o^{j,t-i}, p_a^{j,t-i}, A_o^{j,t-i}, A_a^{j,t-i}; Z^{j,t-i})$$

$$i = 1, 2, \ldots, \infty$$

where: R^{jt} is ridership from location j in period t.

p_o^{jt} is the price of a trip on the train from j in period t.

p_a^{jt} is the price of a trip in the car from j in period t.

A_o^{jt} is the vector of service attributes of the train from j in t.

A_a^{jt} is the vector of service attributes of the car from j in t.

D^{jt} is a distribution of the characteristics of the population in location j at time t which includes the number of people, their destinations, their investments in transportation alternatives, their income, and preferences.

Z^{jt} is a vector of factors unrelated to transportation affecting D.

Substituting for D in the first equation, ridership simply becomes a function of current and lagged price and service attributes and a vector of factors unrelated to previous price and service levels which affect ridership only through demographics.

Several assumptions are made to estimate this dynamic model. These are:

(1) The effects of lagged variables decline geometrically over time.
(2) The decay rate is the same for each explanatory variable.
(3) Attributes of car travel, except for price, are unchanged during the sample period.
(4) Demographic differences across locations (which may be the result of differing prior levels and price of transportation) which give rise to different mean ridership levels across locations can be adequately represented by "fixed effects," meaning that we can use dummy variables for each location.
(5) Z^{jt} is uncorrelated with the transportation variables and so can be disregarded when estimating the impacts of price and service changes.

Given these assumptions, the following equation can be estimated:

$$R^{jt} = \lambda R^{j,\,t-1} + \alpha_1 p_o^{jt} + \alpha_2 p_a^{jt} + \alpha_3 A_o^{jt} + \delta D + e$$

where λ is the geometric lag parameter and D is a vector of dummy variables for location. In the actual regression, dummy variables for years 1984 and 1985 which immediately followed a 3-month strike by SEPTA workers are included as well. The α's give the short-run impact of the variable, while the term $\alpha/(1-\lambda)$ gives the total impact. The mean lag is $\lambda/(1-\lambda)$. The model has been estimated with an asymptotic equivalent of maximum likelihood.

The data set consists of data on 129 of the 165 stations served by SEPTA for twelve observation periods between 1978 and 1986. In addition, the cost of operating, owning and parking a car have been added to the SEPTA data.

The estimation results are presented below. The prices used in the estimation are nominal. The results all conform to what is expected theoretically, and generally the estimated coefficients are highly significant, including those on the lag parameter. The "truncation parameter" is a necessary artifact of the maximum likelihood estimation used here.

Estimation Results: Full Sample, Nominal Prices

Independent Variable	Coefficient	Standard Error	1986 Mean
Lagged Ridership	0.63	0.04	287.2
Peak Number of Trains	5.8	1.6	7.6
Off-peak Number of Trains	3.1	0.6	18.9
Price	-41.8	5.3	5.4
Speed	3.7	1.0	22.0
Variable Cost of Auto Trip	48.8	7.9	8.1
Fixed Cost of Auto Ownership	12.9	6.3	7.1
1984	-81.8	8.7	0.083
1985	-32.5	9.0	0.083
Truncation Parameter	65.8	14.0	

Number of Observations	1548
Mean Square Error	7945.2
Mean Lag	1.7

NOTE: The average time period is eight months, so the mean lag of 1.7 can be converted to 13.6 months. The dependent variable is ridership per station; its mean for the whole sample is 349. The mean for 1986 was 333.

Business Review Index 1987

FEDERAL
RESERVE BANK OF
PHILADELPHIA

BUSINESS REVIEW Ten Independence Mall, Philadelphia, PA 19106-1574