THE WELFARE COST OF RATIONING-BY-QUEUING ACROSS MARKETS: THEORY AND ESTIMATES FROM THE U. S. GASOLINE CRISES*

H. E. FRECH III AND WILLIAM C. LEE

Governments sometimes impose price controls and nonprice rationing-by-queuing. Profit-seeking firms occasionally ration by putting their customers on "allocation." Following Barzel [1974] and Deacon and Sonstelie [1985], we take the decision to ration as a given and analyze it, employing standard microeconomics and applied welfare economics. This paper adds to the literature by focusing on optimally rationing a good *across markets*. Further, we estimate the actual welfare cost of improper allocation across markets in the U. S. gasoline crises of 1973–1974 and 1979.

I. INTRODUCTION

Goverments sometimes impose price controls and nonprice rationing-by-queuing. Generally, they do this against the advice of their economists, as when Richard Nixon imposed price controls in 1971 against the advice of George Schultz, Paul McCracken, and Sam Peltzman, to name only some of the economists in the administration at that time. How are we economists to analyze this behavior?

One view would recommend for us simply to repeat the advice of most economists; namely, that nonprice rationing is wasteful and inefficient, especially when accomplished by waiting.

An alternative view would accept the decision of the government to impose nonprice rationing and proceed to analyze it, employing the usual techniques of microeconomics and applied welfare economics. This is the viewpoint taken by much of the literature on rationing (e.g., Barzel, [1974]; Deacon and Sonstelie [1985a, 1985b]).

This paper adds to the existing literature by analyzing how to ration a good *across markets,* while doing the least harm to consumer welfare. What is more, we provide empirical estimates from U. S. gasoline crises of 1973–1974 and 1979 of the extra welfare losses caused by misallocation of gasoline between urban and rural markets.

*Thanks are due to the Academic Senate Committee on Research of the University of California, Santa Barbara, for financial support. Excellent comments were received from Robert Deacon, Jon Sonstelie, Charles Stuart, Susan Woodward, Larry Summers, and an anonymous referee.

The Quarterly Journal of Economics, February 1987

A queue will be in equilibrium when the implicit cost of a commodity (computed as the value of time spent in the queue plus the explicit money price being charged) equals the price necessary to reduce quantity demanded to the level of the quantity supplied. Diagrammatically, this resembles the imposition of a tax equal to the value of the time spent waiting in line.

However, as with other rent-seeking behavior, the analogy to taxes fails. The quantity that measures the transfer from the customer to the tax collector in the bona fide tax becomes, in the rationing-by-queuing situation, pure waste. The tax analogy similarly fails when we consider the optimal pattern of rationing-by-queuing across commodities or across several markets for a given commodity. In particular, the standard optimal taxation result (for the welfare maximum subject to a revenue constraint), that imposing the highest taxes on the most inelastically demanded commodities is best, must be turned on its head. Instead, for the rationing-by-queuing problem the implicit time price (analogous to the tax) should be greatest for the most elastically demanded good.

In what follows, we first analyze the welfare loss of rationing-by-queuing for a single good sold in a single market. We then develop, as a theoretical benchmark, the optimum rationing scheme for a good sold in several markets with different demand elasticities. Finally, we apply the results of our analysis to the actual allocation between urban and rural travel during U. S. gasoline crises.

II. The Welfare Loss of Rationing-by-Queuing: A Diagrammatic Illustration

Consider the following simplified situation, similar to Barzel [1974]. Consumers are identical; there is no uncertainty; and queue length is constant. The good is being sold continuously in a market with constant demand and supply conditions.

Assume that initially, money price, P_G in Figure I, clears the market so that the representative individual demands Q_G. Now suppose that the government reduces the supply available to the representative consumer to Q_T and decides to ration it by waiting time.

For expositional convenience, we assume income effects to be negligible. Thus, to measure the welfare loss of rationing-by-queuing, we need only to look at a representative consumer's demand curve, illustrated as curve DC in Figure I. The higher total

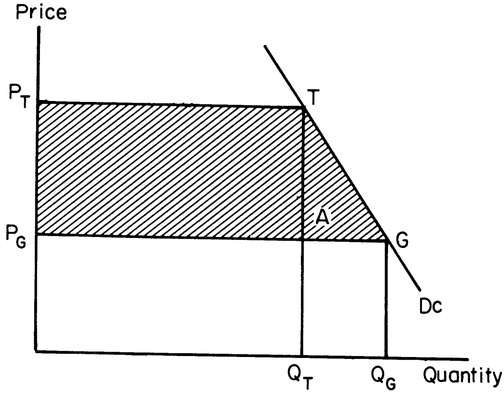

FIGURE I
The Welfare Loss of Rationing-by-Queuing for a Representative Individual

price after the imposition of the price ceiling, P_G, and the nonprice rationing costs the consumer an amount represented by the trapezoid P_GP_TTG. Of this amount, the rectangle P_TP_GTA represents the money value of the time wasted in queuing, while triangle TAG represents the lost surplus simply due to higher total price of P_T over P_G. One may usefully think of TAG as the amount of consumer welfare that would have been lost had the government simply raised the price from P_G to P_T by imposing a tax and redistributed the income.

III. Optimal Rationing-by-Queuing in Many Markets

We assume that the political authority has made a prior decision or is constrained not to allow the price to rise to clear the markets in certain goods (or in certain markets where a good is sold), but that it wishes to minimize losses subject to that decision or constraint. This analysis is designed to find the benchmark that shows the least possible welfare damage done by a policy of rationing by queuing. With this benchmark established, one may then examine how much additional loss is caused by the misallocation of the good across markets—an additional cost that has been largely ignored in the literature.[1]

1. Barzel [1974] stressed the single market, homogeneous consumers' case. The case of a single market with heterogeneous consumers has been examined by Nichols, Smolensky, and Tideman [1971] and Weitzman [1977] in terms of redistributive policy and by Deacon and Sonstelie [1985a, 1985b] in terms of varying values of time.

For analytical simplicity, we assume that the political authority has a limited quantity of one good sold with multiple markets to allocate. In attempting to maximize welfare in the affected markets, the political authority is subject to three constraints. First, it has a limited quantity of the good to allocate—less than the amount necessary to allow price rationing at the fixed price in every market. Second, quantity supplied to each market cannot exceed what is demanded at the fixed money price. Third, the quantities in each market must be nonnegative.

Equivalently, one may think of a multigood problem with a linear production technology. While demand elasticities vary across markets, we assume that consumers within each market are homogeneous. This implies complete dissipation of potential rents by waiting time [Deacon and Sonstelie, 1985b].[2] For simplicity and to highlight our results, we ignore the off-diagonal terms of the Slutsky matrix that might connect the demand in the different markets. And we ignore income effects.[3] Minimizing the losses imposed by rationing is equivalent to maximizing the consumer welfare, subject to the constraints implied by the required money prices and the quantity available for sale. Thus, the political authority faces the nonlinear programming problem of maximizing over P_{Tl}:

$$CV = \sum_{i=1}^{k} \int_{P_{Ti}}^{\infty} Q_i\,(P_{Ti})\,dP_{Ti}, \tag{1}$$

subject to

$$Q - \sum_{i=1}^{k} Q_i\,(P_{Ti}) \geq 0, \tag{1a}$$

$$Q_i\,(P_{Mi}) - Q_i\,(P_{Ti}) \geq 0, \tag{1b}$$

$$Q_i\,(P_{Ti}) \geq 0, \tag{1c}$$

2. Even with heterogeneous consumers, Deacon and Sonstelie [1985b] have shown that rent dissipation can be complete or even more than 100 percent. Suppose that the relatively inelastic demanders have a relatively high value of time. Then the welfare losses exceed 100 percent of the possible rents created by the price control. The assumption of complete dissipation seems reasonable empirically. In an analysis of individual data, Deacon and Sonstelie found waiting time losses to account for 98 percent of possible rents transferred from suppliers to consumers. Of course, if the percentage of dissipation were to vary across markets, the correct allocation across markets should reflect this. And if actual dissipation were to be incomplete, our welfare loss estimates would overstate the losses.

3. Here income effects complicate the optimization as well as estimation of welfare losses. While the consumer's losses can be easily stated in terms of compensating variation, the quantities actually sold, which enter into the constraints, depend on the uncompensated demand functions.

where

$CV \equiv$ welfare, measured as the sum of compensating variations in different markets,
$k \equiv$ the number of markets,
$Q_i \equiv$ quantity demanded in the ith market, as a function of the total price (money price plus time price) in that sector,
$P_{Ti} \equiv$ the total price of the good in the ith sector, including waiting time and money price,
$P_{Mi} \equiv$ the money price of the good in the ith sector, fixed for the analysis,
$Q \equiv$ the total quantity available for allocation to all markets.

These constraints correspond to the requirement that the political authority not distribute more of the good than exists, that the total price must not be below the market clearing price, nor above the choke price, respectively.

The Lagrangian is

$$(2)\quad L = \sum_{i=1}^{k} \int_{P_{Ti}}^{\infty} Q_i\,(P_{Ti})\,dP_{Ti} + \lambda\left[Q - \sum_{i=1}^{k} Q_i\,(P_{Ti})\right] + \sum_{i=1}^{k} \mu_i\,[Q_i\,(P_{Mi}) - Q_i\,(P_{Ti})] + \sum_{i=1}^{k} \pi_i Q_i\,(P_{Ti}).$$

Since P_{Ti} is bounded away from zero by one of the constraints and it will always pay to distribute the total available quantity, the Kuhn-Tucker conditions of interest, ignoring second-order conditions, for our problem are

$$(2a)\quad \frac{\partial L}{\partial P_{Ti}} = -Q_i\,(P_{Ti}) - \lambda \frac{\partial Q_i}{\partial P_{Ti}} - \sum_{i=1}^{k} \mu_i \frac{\partial Q_i}{\partial P_{Ti}} + \sum_{i=1}^{k} \pi_i \frac{\partial Q_i}{\partial P_{Ti}} = 0, \qquad \text{for all } i;$$

$$(2b)\quad \frac{\partial L}{\partial \mu_i} = Q_i\,(P_{Mi}) - Q_i\,(P_{Ti}) \geq 0, \qquad \text{for all } i;$$

$$(2c)\quad \mu_i \frac{\partial L}{\partial \mu_i} = \mu_i\,[Q_i\,(P_{Mi}) - Q_i\,(P_{Ti})] = 0, \qquad \text{for all } i;$$

$$(2d)\quad \frac{\partial L}{\partial \pi_i} = Q_i\,(P_{Ti}) \geq 0, \qquad \text{for all } i;$$

$$(2e)\quad \pi_i \frac{\partial L}{\partial \pi_i} = \pi_i Q_i\,(P_{Ti}) = 0, \qquad \text{for all } i.$$

As is often the case, the interior solution is the easiest to analyze and contains much of the economics. However, the corner solutions are of economic importance, since they are likely to occur. As we shall see, optimal nonprice rationing may require that the money price clear some markets, while at the other extreme the good may have to be removed altogether from other markets. (Constraints (1b) and (1c), respectively, may be binding.) This contrasts with the optimal taxation result that all goods should be taxed.

A. Interior Solution

An interior solution means that some of the good is supplied to each market but not as much as to eliminate nonprice rationing in any market. If so, μ and ρ are equal to zero in all markets. Therefore, equation (2a) simplifies to

$$(2a') \qquad \frac{\partial L}{\partial P_{Ti}} = -Q_i\,(P_{Ti}) - \lambda\,\frac{\partial Q_i}{\partial P_{Ti}} = 0, \qquad \text{for all } i.$$

Considering any two markets and rearranging, we get

$$\frac{-Q_i\,(P_{Ti})}{\partial Q_i/\partial P_{Ti}} = \lambda = \frac{-Q_j\,(P_{Tj})}{\partial Q_j/\partial P_{Ti}}.$$

This form has a nice economic interpretation. The numerator is the marginal loss in consumer welfare as the total price is raised. The denominator is the marginal reduction in quantity. The ratio gives the marginal cost of reducing quantity in one particular market. The equation holds that this marginal cost must be equal in all markets. In terms of the absolute values of the own price elasticity of demand, η_i, this condition can be written as

$$(3) \qquad P_{Ti}/|\eta_i| = P_{Tj}/|\eta_j| = \lambda.$$

That is, the total price, including waiting time, is directly proportional to the elasticity. Optimality requires the highest total price in the market with most elastic demand. Although at first blush surprising, the economics of this result can easily be seen graphically. Suppose that total quantity must be cut by one unit below the equilibrium shown in Figure II. Which market should the one unit be taken from? If it is taken from the elastic market, welfare losses (ignoring income effects) will be given by the shaded area. If the unit is taken from the inelastic use, the losses would be the much larger cross-hatched area. Taking the unit from the elastically demanded sector results in a lower loss because less of the wasteful waiting time is necessary to reduce demand enough for markets to

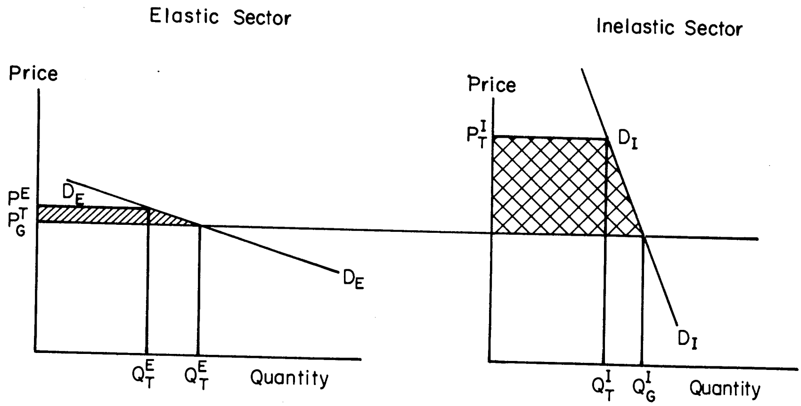

FIGURE II
Optimal Nonprice Rationing Across Sectors

clear. Of course, positive cross elasticities of demand would reduce the optimal total price differences.

Further, an analogy to a price-discriminating monopolist may be helpful. The rationing process is equivalent to one in which a monopolist is collecting rent from different consumer groups by charging different prices in different markets and then wasting the rent. Therefore, a waste-minimizing policy would reverse the textbook price discrimination rule. It would charge lower prices to consumers with less elastic demands, which is the policy derived here.

B. Corner Solutions

Several interesting corner solutions can arise. First, for small quantity cutbacks from money price market clearing, it may be optimal for some markets to bear the entire quantity reduction, while there is no quantity reduction, and hence no nonprice rationing, in other sectors. Algebraically, this occurs when constraint (2b) is binding for some (not all) i. Therefore, some $\mu_i > 0$, while some $\mu_j = 0$. To highlight the economics of the corner solutions one by one, let us assume that constant (2c) is not binding so that $\pi_i = 0$ for all i. This means that some of the good is sold in every market. We shall examine the possible optimality of zero supply to some markets below.

Let us select a market 1, which is not rationed and a market 2, which is. The first-order conditions can be expressed in terms of absolute values as

$$P_1/|\eta_1| > P_2/|\eta_2|. \tag{4}$$

This type of corner optimum is reasonably likely, especially for relatively small quantity reductions from price rationing equilibria. For example, suppose that we have constant elasticity demand curves. If one is twice as elastic as the other, that more elastic use should bear the full brunt of the cutbacks until the total price is twice that of the unrationed use. Only at that point should the less elastic use be reduced at all. Note that if there is a choke price (where none of the good is demanded), near that choke price the elasticity rises faster than price, so that the authority should cut back one quantity all the way to zero in the most elastic market before starting nonprice rationing in another market. This brings us to our next case of corner solutions.

If some markets are cut back so severely as to be eliminated, optimality requires that constraint (2c) is binding for some (not all) i. Therefore, some $\pi_i > 0$, while some $\pi_j = 0$. Parallel to the above, let us assume that there is some quantity reduction in every market so that $\mu_i = 0$, for all i.

Let us select a use, 3, which is cut back to zero and a use, 4, which is at an interior position (nonprice rationed). The first-order conditions, after rearranging analogously to the above case are

$$P_3/|\eta_3| < P_4/|\eta_4|. \tag{5}$$

The political authority would like to cut back further on use 3, the elastically demanded one. However, it has already cut back to zero! Totally eliminating use in one or more markets can be efficient. In fact, this was the main nonprice rationing scheme in World War II America. Production of consumer durables was made illegal [Tobin, 1964].

IV. An Application: The Gasoline Crises of 1973 and 1979

In 1973 and again in 1979, the combination of a reduced supply of crude oil from the Mideast and U. S. price controls produced long lines of cars queuing up for gasoline. Retail price controls were binding. The political authority had decided that, for a time, money price would not be allowed to clear the market. Given this decision, there were more and less efficient ways to allocate gasoline across markets. In this section we empirically apply the analysis above to the problem of allocation between the rural and urban markets within the state of California. We first determine the optimal allocation of gasoline and associated total prices, and this forms the benchmark against which to compare actual allocations, implicit prices, and welfare losses.

We require estimates of demand by market to apply the analysis. Traditional gasoline demand studies use data on the gallons of gasoline sold as the quantity measure. Since this measure is not available for separate gasoline markets, these studies will not do for our purpose. However, Lee [1976, 1980a, 1980b] has estimated demand functions separately for different types of actual trips in California over the decade of the 1970s, using monthly data. The dependent variable is the number of vehicles passing over several traffic counters maintained at various locations classified as urban or rural across the state by the California Transportation Department.

The estimates of demand elasticities are -0.189 and -0.305 for the 1970–1975 rural and urban travel and -0.289 and -0.342 for 1976–1979 rural and urban trips.

Demand was more elastic for urban than rural travel during both of the episodes of gasoline rationing-by-queuing.[4] This higher elasticity for urban travel requires that urban travel should be nonprice rationed first, and perhaps urban travel alone should be rationed. To see whether, in fact, urban travel alone would optimally bear the entire burden of nonprice rationing, one must examine the ratio of total prices to the absolute value of elasticities in the two markets. For the cases most likely to lead to interior optima, the largest cutbacks, this leads to

$$(7)\qquad \left(\frac{P_{RUT}}{|\eta_{RUT}|} = 6.06\right) > \left(\frac{P_{URT}}{|\eta_{URT}} = 1.51\right), \qquad \text{for March 1974;}$$

$$(8)\qquad \left(\frac{P_{RUT}}{|\eta_{RUT}|} = 4.19\right) > \left(\frac{P_{URT}}{|\eta_{URT}|} = 2.03\right), \qquad \text{for May 1974.}$$

Compare with equation (3). Thus, optimality requires rationing-by-queuing exclusively in urban transportation. In fact, naturally arising uncertainties about supplies, short hours, and government-mandated Sunday closings led to far more severe rationing for rural

4. This may seem surprising. However, cities offer many alternative means of transport (public transportation, carpooling, even walking). For rural trips, there is often no substitute for automobile travel. In urban areas, work trips are accomplished by public transport for 8.2 percent of the workforce, while in rural areas public transport is used by only 0.7 percent of workers [*Statistical Abstract*, 1979, p. 653].

TABLE I
MONEY PRICES, TIME PRICES AND GASOLINE CONSUMPTION IN CALIFORNIA

1973–1974 Crisis		Dec. 1973	Jan. 1974	Feb. 1974	Mar. 1974
Money price (1985 $/gal.)		1.009	1.044	1.047	1.148
Waiting time price (1985 $/gal.)	URBAN	0.126	0.135	0.445	0.570
	RURAL	0.242	0.229	0.445	0.964
Gallons per day (×1,000)	URBAN	5,493	5,091	5,029	4,850
	RURAL	17,329	15,304	16,914	15,525

1979 Crisis				
Money price (1985 $/gal.)		1.215	1.263	1.325
Waiting time price (1985 $/gal.)	URBAN	0.403	0.126	0.081
	RURAL	0.380	0.345	0.129
Gallons per day (×1,000)	URBAN	7,095	7,894	8,132
	RURAL	22,783	24,867	26,402

travel, as reflected in the higher implicit waiting time price for gasoline for rural use in Table I.[5] This was especially true for the 1973–1974 crisis, during which the government took more active steps to discourage rural travel [*Wall Street Journal*, 1973, 1974]. For example, in March 1974 the rural time price was $0.96 per gallon; while the urban one was only $0.57 per gallon.

A. Welfare Losses

Table II gives the welfare losses actually incurred, with greater rationing for the rural market versus the minimal losses that would have occurred if only the more elastic urban market had been rationed. Since income effects are small, the ordinary demand curve is used to approximate the compensated one.[6] As one can see, the extra loss due to incorrect allocation across markets is quite large.

V. CONCLUSION

Optimal nonprice rationing-by-queuing requires that the most elastically demanded good should be the most severely rationed. Further, optimality can require a decidedly nonmarginalist

5. Tables I and II exclude crisis months for which we do not have good estimates separately for urban and rural implicit prices. See Lee [1976, 1980a, 1980b].

6. Gallons were allocated to urban or rural travel based on the estimated ratio of gallons per rural trip to gallons per urban trip, 2.34.

TABLE II
WELFARE LOSS OF RATIONING-BY-QUEUING IN CALIFORNIA
(IN THOUSANDS OF 1985 DOLLARS)

1973–1974 Crisis					
	Total period	Dec. 1973	Jan. 1974	Feb. 1974	Mar. 1974
ACTUAL					
Urban	$ 189,068	$ 21,781	$ 21,781	$ 67,503	$ 78,003
Rural	986,941	133,284	115,100	221,008	517,549
Total	$1,176,008	$155,065	$136,881	$288,510	$595,552
OPTIMAL					
Urban	$670,500	$99,214	$91,820	$191,137	$288,349
Rural	0	0	0	0	0
Total	$670,500	$99,214	$91,820	$191,137	$288,349
Excess	$505,508	$55,851	$45,061	$ 97,373	$307,223

1979 Crisis				
	Total period	May 1973	June 1979	July 1979
ACTUAL				
Urban	$144,165	$ 94,218	$ 30,264	$ 19,683
Rural	653,486	282,154	268,121	103,210
Total	$798,618	$376,372	$298,385	$122,893
OPTIMAL				
Urban	$593,332	$296,542	$199,375	$97,415
Rural	0	0	0	0
Total	$593,322	$296,542	$199,375	$97,415
Excess	$204,319	$ 79,830	$ 99,011	$25,478

approach, in that the elastically demanded use or good takes the full brunt of the nonprice rationing.

We apply our welfare results to rationing-by-queuing during the U. S. gasoline crises of 1973–1974 and 1979, for which we separately estimate the demand for gasoline in urban and rural markets within California. We find that the government reduced the supply most drastically in the most inelastically demanded market—that for rural travel. This error of allocation across markets raised welfare costs above what they would have been with optimal rationing-by-queuing by about 75 percent ($506 million in California alone) in the first crisis and by about 34 percent ($204 million in California alone) in the second.

The applicability of this analysis is wider than our focus on rationing-by-queuing would suggest. It would apply to any nonprice rationing where the market is cleared by a resource-using rent-seeking process such as political competition [Tullock, 1967; Comanor and Leibenstein, 1969; Posner, 1975] the gathering of privately valuable but socially useless information [Hirshleifer, 1971; Marshall, 1974], or nonprice competition [Frech and Samprone, 1980]. In any such cases, optimal allocation across rationed goods or markets for a good would have to meet rules much like ours.

UNIVERSITY OF CALIFORNIA, SANTA BARBARA
ST. MARY'S COLLEGE

REFERENCES

Barzel, Yoram, "A Theory of Rationing by Waiting," *Journal of Law and Economics,* XVII (April 1974), 73–96.

Comanor, William S., and Harvey Leibenstein, "Allocative Efficiency, X-Efficiency and the Measurement of Welfare Losses, "*Economica,* XXXI, N.S. (August 1969), 304–09.

Deacon, Robert T., and Jon Sonstelie, "Rationing by Waiting and the Value of Time: Results from a Natural Experiment," *Journal of Political Economy,* XCIII (August 1985a), 627–47.

——, and ——, "The Welfare Costs of Rationing by Waiting," unpublished paper, Economics Department, University of California, Santa Barbara, October 1985b.

Frech, H. E. III, and Joseph C. Samprone, Jr., "The Welfare Loss of Excess Nonprice Competition: The Case of the Property-Liability Insurance Industry," *Journal of Law and Economics,* XXI (October 1980), 429–40.

Hirshleifer, Jack, "The Private and Social Value of Information and the Reward to Inventive Activity," *American Economic Review,* LXI (September 1971), 561–73.

Lee, William C., "The Demand for Travel and the Gasoline Crisis," Ph.D. thesis, University of California, Santa Barbara, 1976.

——, "Demand for Travel and the Gasoline Crisis," *Transportation Research Record,* No. 764 (1980a), 38–42.

——, "The Demand for Travel and the Gasoline Crises," Final Report to the Motor Vehicle Manufacturers Association, June 1980b.

Marshall, John M., "Private Incentives and Public Information," *American Economic Review,* LXIV (June 1974), 373–90.

Nichols, D., E. Smolensky, and T. N. Tideman, "Discrimination by Waiting Time in Merit Goods," *American Economic Review,* LXI (June 1971), 312–23.

Posner, Richard A., "The Social Costs of Monopoly and Regulation," *Journal of Political Economy,* LXXXIII (August 1975), 807–27.

Tobin, James, "A Survey of the Theory of Rationing," *Economica,* XX (October 1964), 540–48.

Tullock, Gordon, "The Welfare Costs of Tariffs, Monopolies, and Theft," *Western Economic Journal,* V (June 1967), 224–32.

Statistical Abstract of the United States (Washington, D.C.: U. S. Government Printing Office, 1979).

Wall Street Journal, "President Nixon Requests Sunday Ban on Retail Gasoline Sales" (November 2, 1973), p. 3.

——, "Gasoline Station Sunday Closings Ended by President Nixon" (March 20, 1974), p. 3.

Weitzman, Martin L., "Is the Price System or Rationing More Effective in Getting a Commodity to Those Who Need It Most," *Bell Journal,* VIII (Autumn 1977), 517–25.