A Theory of Rationing by Waiting

Yoram Barzel

Journal of Law and Economics, Vol. 17, No. 1. (Apr., 1974), pp. 73-95.

Stable URL:
http://links.jstor.org/sici?sici=0022-2186%28197404%2917%3A1%3C73%3AATORBW%3E2.0.CO%3B2-X

Journal of Law and Economics is currently published by The University of Chicago Press.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ucpress.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org/
Tue Apr 4 15:10:43 2006

A THEORY OF RATIONING BY WAITING*

YORAM BARZEL
University of Washington

WHEN limited quantities of commodities are provided free by the government, it is frequently asserted that not all demand is satisfied. Such commodities are often allocated on a first-come-first-served basis. It will be shown that rationing by waiting, which is the economic counterpart to first-come-first-served, is just another way to "satisfy all demand." The mechanics and the implications of such rationing are the subject matter here.[1]

Compared with the price mechanism, rationing by waiting generally entails a different resource cost—the time spent in the queue.[2] The extent of that cost, sometimes designated "dissipation of rent," will be examined and the conditions specified under which it equals the entire consumer's surplus. The cost arises because no well-defined property rights originally exist to the "free" good. To establish a claim to such rights, one must spend whatever time is necessary to gain priority in the queue.

The provision of goods free of charge is often justified on grounds that the action will benefit the poor. However, it will be shown that even when the poor actually have a lower time cost and therefore a lower cost of waiting, they may not benefit from such action. Indeed, in some empirically relevant cases the rich rather than the poor will stand in line to obtain such goods.

Assuming that every individual is costlessly informed of the waiting time necessary to acquire a subsidized good, what can we say about the length of

* I wish to thank Mary Eysenbach, Chris Hall, Reuben Kessel, John McGee, Mike Rahm, and particularly Steve Cheung for their valuable comments. I was first exposed to the notion of the real costs induced by price control in Harberger's price theory course in Chicago in 1957. Thanks for secretarial help are given to the National Science Foundation Grant to the University of Washington for the investigation of contracts.

[1] The literature on rationing is extensive, but with few exceptions the rationed quantities are assumed to be allocated by mechanisms such as a point system, for which this kind of waiting does not arise. Some queuing problems are discussed by D. Nichols, E. Smolensky, & T. N. Tideman, Discrimination by Waiting Time in Merit Goods, 61 Am. Econ. Rev. 312 (1971) [hereinafter cited as Nichols]. Gary S. Becker discusses briefly the cost associated with queues in A Theory of the Allocation of Time, 75 Econ. J. 493 (1965).

[2] Queues occur also in unregulated markets. While some of the analysis of this paper may apply to such situations, we concentrate here on queuing problems emerging from government actions.

the queue or, more precisely, about the length of waiting time for each individual in line? For this type of market to clear, the *marginal* individual is at a point where *total cost* (time plus money) is equal with his *total* valuation of the good to be obtained. In other words, to the marginal individual the consumer's surplus is zero.

A recent event, verging on a controlled experiment, confirms that allocation by waiting is not just an economist's fancy but is actually a market phenomenon. On June 14, 1972, the United States of America Bank (of Chicago) launched an anniversary sale.[3] The commodity on sale was money, and each of the first 35 persons could "buy" a $100 bill for $80 in cash. Those farther down the queue could each obtain similar but declining bonuses: the next 50 could gain $10 each; 75, $4 each; 100, $2 each, and the following 100, $1 each. Each of the next 100 persons could get a $2 bill for $1.60 and, finally, 800 (subsequently, it seems, expanded to 1800) persons could gain fifty cents each. The expected waiting time in such an unusual event was unpredictable; on the other hand, it was easy to assess the money value of the commodity being distributed.

First in line were four black brothers aged 16, 17, 19, and 24. Since the smallest was 6′2″, their priority right was assured. "I figured," said Carl, the youngest brother, 'that we spent 17 hours to make a $20 profit. That's about $1.20 an hour."

"You can make better than that washing dishes," added another of the brothers. Had they been better informed they could have waited less time. The 35th person to join the line arrived around midnight, had to wait just 9 hours, and was the last to earn $20—$2.22 per hour. To confirm her right, she made a list of all those ahead of her in the line.

"Why am I here?" she asked. "Well, that $20 is the same as a day's pay to me. And I don't even have to declare it on my income tax. It's a gift, isn't it?"

Information is sparse on how the rest of the queue evolved, but by opening time an estimated 500 persons were in line. Between then and the noon deadline, another 1800 or so joined the queue to get fifty cents apiece. What emerges most clearly from this experience is that the identity of those in the queue was not random and that waiting time fell as the bonus declined. The money value of the "free" good was largely matched by waiting time, viewed as a cost by those in the queue.[4]

[3] The information is obtained from the Seattle Times, June 12, 1972, and from four Chicago daily newspapers, the Chicago Sun Times and Chicago Today, June 14, 1972, and the Chicago Tribune and Chicago Daily News, June 15, 1972.

[4] From the viewpoint of the bank, which apparently sought publicity, waiting time was simply a purchased input; indeed, it turned out to be most productive, as attested by the prominent newspapers' writeups.

Time as a Market Clearing Price

The term "rationing" is traditionally identified with disequilibrium, a condition under which individuals are unable to equate costs and values on the margin, as seems apparent when the price imposed is not at the intersection of demand and supply. But since time is costly, waiting provides an additional route whereby individuals can again equate on the margin. Adding the time constraint allows us to apply "equilibrium" analysis to this form of rationing.

In Figure I the horizontal axis measures the quantity of a commodity and the vertical axis its money price. The conventional demand curve of a single individual is represented by $d(t_0)$. (Implicit is the notion that the pecuniary price represents the full cost of purchase where waiting time is zero, that is, $t_0 = 0$.) For each of the other curves, waiting time is assumed to be positive and, at the higher time price, these curves have to lie below $d(t_0)$. At $d(t_2)$ waiting time per unit of the commodity is $t_2 > 0$; at $d(t_3)$ waiting time is $t_3 > t_2$.

When the money price is reduced to zero and per-unit waiting time is t_3, the quantity demanded is k. If k is the per-person quantity being distributed "free" by the government, the individual will be willing to stay in line up to $k \times t_3$ minutes to obtain his share.[5] By letting t vary continuously, every point inside $d(t_0)$ is a point on a demand curve. In this space, for every combination of quantity and money price there is a time price at which the individual will be able to equate on the margin and be in equilibrium. In particular, all points on the horizontal axis to the left of the $d(t_0)$ horizontal intercept, combined with the appropriate amounts of waiting, are equilibrium points. Holding the pecuniary price constant, we can trace the relation between the quantity demanded and the minutes-per-unit waiting price to obtain a demand curve in terms of waiting, as we will do below for the special case where the monetary price is set at zero.

It is important to recognize that when money price is zero, waiting time serves as a total price, and the analogy with pecuniary price is complete: demand with respect to this price will slope downwards. In the absence of discrimination among consumers only one time-price will prevail in the market.[6]

[5] If it is impossible to get fewer than k units at a proportionately lower waiting time, all-or-nothing demand becomes relevant. Such a demand relation is introduced *infra*.

[6] The conclusion that waiting time is constant is derived under the assumption of certainty. P. Naor, The Regulation of Queue Size by Levying Tolls, 37 Econometrica 15 (1969) shows that with a random stream of customers the length of the queue will vary but will never exceed a given level equivalent to our uniform waiting time. He implicitly assumes there is no cost in ascertaining the length of the queue, so that at its maximum level there is no incentive to join. Symmetry would require, then, that when the queue is shorter then maximum, there is an incentive to join thereby lengthening the time in the queue towards the maximum. In any case, when the cost of information is taken into

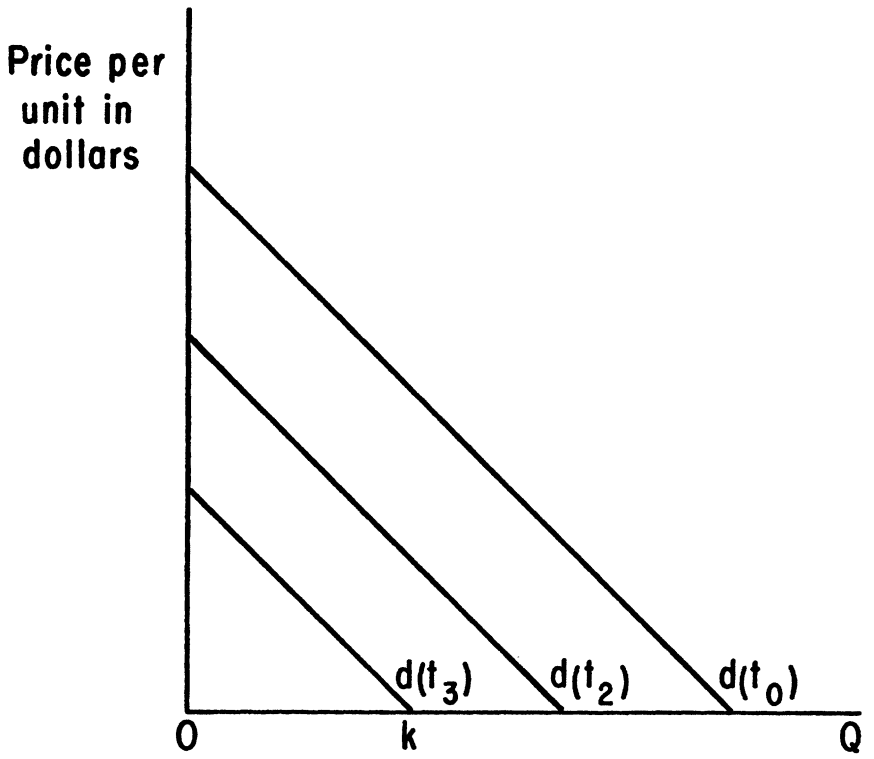

FIGURE I

The marginal valuation here is in terms of waiting time: as the quantity offered increases, the marginal valuation and, consequently, the per-unit waiting time will decline, giving the curve its negative slope.[7]

Unless the individual allotment of the "free" good is restricted, one person may take the entire available supply. Thus the rules for distribution will fix quotas and may also stipulate whether each may join the queue only once

account, the decision whether to join will depend on the probability that the queue is shorter than the critical level. The higher this probability the stronger the incentive to explore, which in turn reduces the variability in the length of the queue.

[7] Given that a certain quantity of a commodity is distributed at a zero price, what can be said with respect to changes in the quantity? The cost of increasing the quantity is positive; but as the quantity increases, so does the rent to consumers since waiting time falls. If one is willing to add up dollar valuation of different individuals, a local maximum will occur when the rate of change in the sum of rents to consumers is equal to the rate of change in total cost of production. Second order as well as total conditions have to be also satisfied. In addition, it seems possible that several local maximum points may exist.

per period or as many times as he wishes. The former arrangement prevails with the food stamp program and housing subsidies; the latter implicitly governs the utilization of public parks and roads. Both techniques are considered here.

The time-price an individual is willing to pay per unit of the good falls when the individual allotment increases. And, holding the per-person quantity constant, the number of demanders will increase as the time-price falls. Just as with regular demand, so these two factors account for the negative slope of the aggregate demand curve when waiting time is the price. It might appear paradoxical that time decreases per unit of the good as more is made available for distribution and as the number waiting in line increases. The length of the queue may grow longer when the number of persons being served is larger. But regardless of the efficiency of distribution, for fixed individual allotments per-person waiting has to decline as the number of allotments is increased.

We will now demonstrate that per-person time in the queue is basically independent of the mechanics of distribution, first for a most efficient and speedy organization and then for a relatively slow one. Suppose that for a given amount of a commodity to be distributed free, one batch to a person, the per-person waiting time needed to clear the market is one hour. Now if the entire available amount can be dispensed instantaneously, the *whole* queue will have to form instantaneously one hour before distribution starts so that the time price is the same to all. If the line does not form instantaneously, the early joiner wastes his time, since he could still have obtained the commodity by coming later. He will do his best to join as late as possible, which is when the marginal man is indifferent between being in or out. Waiting cannot, however, be reduced to less than one hour, since positions will then be preempted by extramarginal individuals.

When the number of batches being distributed is increased, the number of individuals in the queue obviously will be larger, but, since the newcomers value a batch of the commodity less than they value one hour of their time, the market-clearing waiting time is shorter—say half an hour. Individuals will now line up only half an hour before distribution starts, or pressure will be exerted to bring it to that level. Thus the queue is longer, but per-person waiting time is less.

If the distribution arrangement is less efficient so that—say—six minutes are required to give out to one person the same amount as before, the first individual will arrive 54 minutes before distribution commences,[8] and from then on others will join the queue at six-minute intervals so that each

[8] This is so because 54 minutes are spent in line before distribution starts and 6 minutes for actually getting the good.

spends one hour waiting. No other spacing would persist, because if waiting takes more than one hour, some individuals will drop out; if less, some extra-marginal individuals will join. After the process is stabilized, the number in the line will be exactly 10, with accretions matching completers, and that level will hold until the last man appears one hour before closing time. For a larger quantity the waiting time required to clear the market is shorter. The total process will be streched over a longer span (or over more distribution centers) so that the additional unit can be distributed; but at any point in time, except at the two ends, fewer individuals will be in the queue.

In the first situation it may seem that if the distribution were started one hour earlier, the line would disappear; but as the new opening time becomes known—instantaneously under the assumptions here—the queue would simply form yet one hour earlier. In the second situation, it may seem that if only the bureaucrats took fewer coffee breaks, less time would be wasted standing in line. But again the queue would eventually adjust to the new conditions and equilibrium waiting time would be restored to one hour.[9] An important corollary is that, given a fixed amount to be distributed free, attempts to improve efficiency will be wasted, since time in the queue will not change. Conversely, the lowest-cost method of distribution, no matter how cumbersome it may appear, is the most efficient in economic terms, since consumers' valuation is independent of such costs.

The speed of the distributing arrangements, however, may affect waiting time in two minor ways. First, under the terms of our numerical illustration, if the available quantity is so large that the equilibrium waiting time is less than six minutes, "technical" conditions then force the waiting price to six minutes. At that time-price some of the available amount will not be demanded and quicker handling is needed to reduce time below six minutes. Second, the distributing organization may be so small that it can handle only part of the available amount. The analysis may now be carried on as before, with the quantity capable of distribution rather than the quantity available constituting the effective constraint.

Even when individuals are allowed to reenter the queue, the available supply may already have been exhausted by the time the first person is

[9] Gary S. Becker, *supra* note 1, at 516, n.3, states, "Time costs . . . often result in external diseconomies: for example, a person joining a queue would impose costs on subsequent joiners." This may appear contrary to our conclusion that the larger the number of batches to be distributed, the larger the number of waiters and the *lower* the waiting per person. Becker's statement, however, seems to apply to the process by which a queue evolves in the absence of adequate information rather than to differences among equilibria of distinct queuing situations for which a *larger* number of persons going through a queue results in a *shorter* per-person waiting time.

ready to step in line again. The quicker the distribution, the greater is the chance this will happen. Here, then, is a case where the distribution arrangement may significantly affect the market-clearing waiting time: speedier distribution shortens the waiting time, since there is less opportunity for multiple approaches.

Several major implications which, at least in principle, are testable emerge from the above discussion. We argue that time-price plays the same role as money price and that for a given stock of a "free" commodity a market-clearing time-price will be reached. Consequently, as the per-person allotment is increased with the number of persons served held constant, we predict that total per-person time in the queue will increase but that this increase will be less than proportionate to quantity, since time per unit is predicted to fall. In addition, holding the per-person allotment constant, per-person time in the queue will decline as the number of allotments is increased. And finally, since we argue that knowledge of the per-person quantity and of the number of persons served is sufficient to determine the market-clearing price, a testable implication is the invariance of waiting time in the queue with respect to the speed of the distribution.

The Resource Cost of the Queue

In the main part of this section it is implicitly assumed that the only cost of getting a commodity to the final consumer is the cost of its production. Let us look briefly at some of the problems associated with this rather basic assumption. In the case of private markets, costs associated with selling are often considerable. On the other hand, for freely disbursed commodities such as roads, elementary public schools, and parks, no resources are used in rationing the available quantity among consumers. The reduction in selling costs as compared with private markets is real enough. We cannot expect, however, that either the quantity supplied or the quality mix will be the same as those supplied by the market.

In any case, it is possible to avoid explicit rationing of roads, schools, and public parks because the rate at which consumers can acquire services is basically restricted to one unit per unit of time. Moreover, it is often prohibitively costly to resell such commodities. In the case of food, on the other hand, the first man in the queue may acquire the entire stock, either for his own consumption or to sell to others. The cost of a rationing organization to prevent this parallels the cost of selling in private markets. Adequate comparison between the two forms of providing commodities would require full assessment of the two types of rationing costs and would be beyond the scope of this paper, therefore we abstract from this problem, as follows.

What is the nature, and measure, of the cost associated with queuing? The mere presence of non-price rationing suggests that the quantity supplied differs from what would be offered in the absence of government intervention. The "misallocation" or "waste" associated with the government action may be measured, say, by the appropriate triangles. Time spent in the waiting line is an additional cost.[10] A pecuniary market price represents a transfer of resources from the buyer to the seller to the extent of the resource cost of producing the good. When a queue is the means of rationing, the time-resource spent to acquire the good is not captured by the producer. If the quantity provided free is equal to what the market would provide, no misallocation triangle appears, but the resource cost of the queue must still be accounted as additional to that of producing the good. Indeed, the two types of cost are basically the same for the marginal unit and are of the same order of magnitude for the entire quantity.

Suppose it were possible costlessly to identify in advance all those planning to be in the queue and to provide them with coupons to claim the good. The saving of time in the lineup would be a net gain, since no other resources were expended. In the absence of such arrangement, since the marginal man in the queue is indifferent between spending his time there or in some alternative use, the extent of the cost for that individual is identical with his total valuation of the "free" good.[11]

Consider a situation in which a large number of individuals have access to a good to be rationed at zero monetary price by means of a queue. In Figure IIa the demand curve d_i of the i^{th} individual is drawn with price in terms of waiting time. Points on this demand curve are implicit in Figure I and can be obtained by matching the quantity and the per unit time price at the intersection of each demand curve with the horizontal axis where pecuniary price is zero. For instance, as evident in Figure I, the time price the individual is willing to pay for k units is t_3 minutes per unit, which gives the point (t_3, k) on d_i in Figure IIa.

[10] At a deeper level, "misallocation," "waste," and time in the queue may be viewed as costs similar in nature to those of selling and transacting. We might hypothesize that these are among the costs which the economy will attempt to minimize. Thus it is conceivable that the existing arrangement is the lowest in cost for achieving certain objectives, and we will touch on this notion later.

[11] In popular discussion of conditions associated with consumption of free goods, the nature of the cost involved is usually (at best) only implicit. However, to cite one exception: "Of course, public [tennis] courts are free or relatively cheap but not always freely available for playing on. In Houston the public courts have become so overtaxed by the tennis boom that players must appear in person a full week before they want to play in order to sign up for their hour on the court. What the poor or parsimonious player may save in money, he can easily spend in time and inconvenience." Sophy Burnham, Tennis: A Whole New Ball Game, Saturday Review, July 29, 1972, at 44, 46.

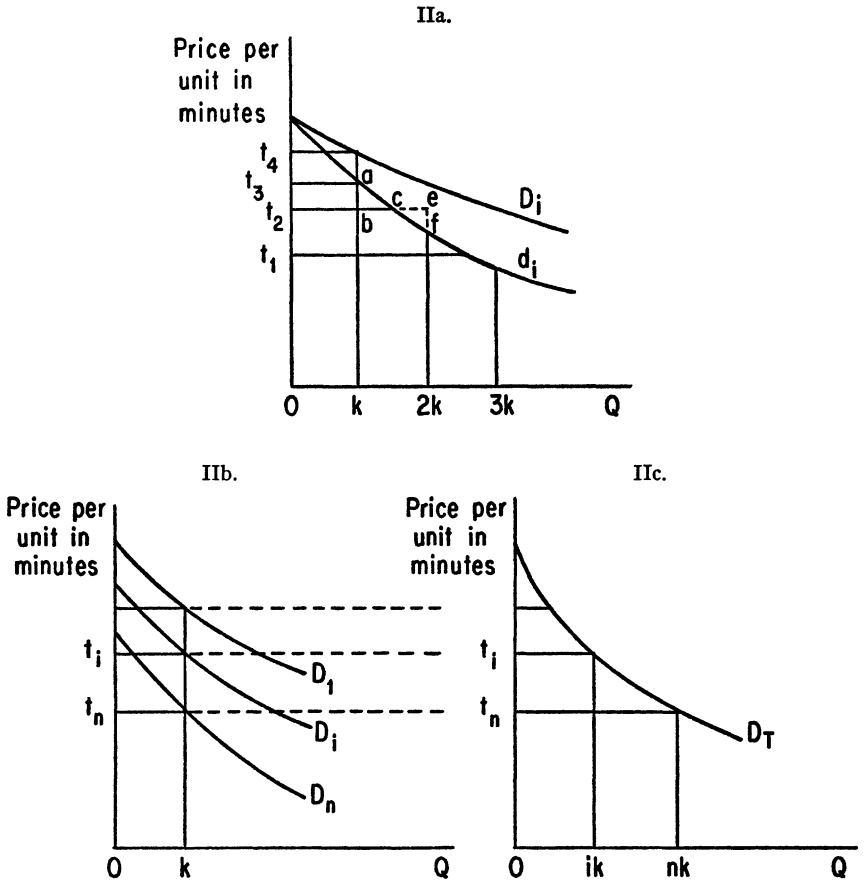

FIGURE II

If the individual is faced with an "all-or-nothing" decision to acquire k units, he will be willing to pay the entire area below the demand curve to quantity k, or an average price t_4, for the quantity.[12] D_i is the all-or-nothing demand curve corresponding to d_i.

In Figure IIb, the all-or-nothing demand curves for several individuals are drawn, ordered (in the relevant range) from highest to lowest. While only a few individuals are represented in the drawing, we assume that the total number is large and that at any quantity the distance between adjacent curves is small enough to be ignored. If the batch size is k, for any time-price

[12] We abstract from the difference between income underlying the all-or-nothing decision and that underlying the regular demand curve.

we obtain the number of batches demanded by simply counting the number of individuals for whom the height of the all-or-nothing demand curve at quantity k is equal to or exceeds the specified time price. Given a fixed supply of n batches, price is market determined, and its equilibrium level is the height of the all-or-nothing curve of the n^{th} individual. At that price, the $(n + 1)^{st}$ individual will not demand the good, since his total valuation of the batch is less than his cost. At any lower price, the $(n + 1)^{st}$ individual is better off with the good than without it; consequently, he will be in the line. The available quantity, however, will not provide the good to all those waiting, and this exerts upward pressure on the time-price, bringing it back to the height of the all-or-nothing demand of the n^{th} individual. Thus the equilibrium, or market-clearing, price for $n \times k$ units is the all-or-nothing price for k units for the n^{th} individual, who therefore earns no surplus since his total valuation is precisely exhausted by waiting.

The aggregate demand curve with batch size k and a single access to the queue is drawn in Figure IIc. Notice that since the per unit market-clearing price for n batches is t_n, total waiting time is $n \times k \times t_n$ which represents the total time spent in queuing.[13]

When more than one entry in the line is permitted, the demand curve becomes a mixture of the regular and the all-or-nothing curves. The all-or-nothing component applies to the last batch, while all previous ones are obtained at the going time-price.[14] At price t_2, the i^{th} individuals will certainly join the queue and demand one batch, as shown in Figure IIa. The height of t_2 determines that the area of the triangle abc equals that of cef, and so the valuation of the second batch equals its cost. When the price is t_2, the i^{th} individual, then, would just be willing to join the queue a second time. Similarly, he will stand in line three times when the price falls to t_1, since t_1 is his all-or-nothing price for the third batch.[15] For an individual obtaining just one batch, the conventional all-or-nothing demand holds. At each price, marginal consumers can be classified into two groups: those who have obtained at least one previous batch, and those just indifferent between getting or not getting one single batch.

[13] In the above discussion, n, k (and their product nk) were treated as given. Provided that some quantity is distributed free, one may wish to explore why the particular values of n and k were selected.

[14] Again we abstract from differential income effects.

[15] This analysis differs from the common textbook analysis of demand in that the units of the commodity being distributed are assumed discrete rather than continuous. If the batch size goes to zero, a regular demand curve results. But in the real world the unit by which commodities are sold is seldom fully divisible, consequently the analysis here is more appropriate than that found in most textbooks.

Now n batches will be distributed among fewer than n individuals since at any time-price we would expect, in general, some consumers to obtain more than one. But some will enter the line only once and to the marginal individual in this group, the "free" good provides zero rent just as in the case of single access to the queue. Total time spent in queuing here is $n \times k \times t'_n$ where t'_n is the market-clearing time-price for multiple access (which will be higher, in general, than t_n). We have shown, then, that whether the rule is single access or multiple access, the entire consumer's surplus that the marginal man in the queue derives from the good is exhausted by waiting.

If all individuals under consideration are identical in their utility functions, incomes, and time costs, then the marginal and the average individual are identical. Since the market-clearing waiting time is such that the marginal person in the line receives zero rent (or surplus), so too does the average person. Each is indifferent between having and not having the good, and the entire valuation of the good is lost in waiting.[16]

If those obtaining the good are not identical, the marginal individual is the only one for whom the entire economic rent is lost, and each of the intramarginal ones receives some rent. The rent is equal to the difference in valuation between the maximum amount of time that the particular individual is willing to wait and the actual waiting time. Given a quantity to be distributed free, the aggregate rent is larger the more diverse are those in the queue in terms of their willingness to wait. Homogeneity leads to highly elastic demand for a good whose price is in waiting time, to low returns to those in the queue, and to relatively high cost of waiting in terms of loss of alternative output. Thus the notion that the cost of the free good is low is not generally correct, and we might reject the statement that "to the low wage people, the money cost of the queue is *minimal* and they will receive a *substantial* benefit due to the lower money price."[17]

Turning the problem around, we gain added insight into the political process involved. Queuing uses resources the return from which, on the average, is larger the lower the demand elasticity of the subsidized commodity. If the demand is perfectly elastic, no consumer gains from the subsidy; consequently, little political support can be expected from consumers for the subsidization of such commodities, the less elastic the demand, the larger the gains on intramarginal units and the greater potential political support. We predict, then, that commodities provided free and distributed on a first-come-first-served basis will have, on the average, lower demand elasticities in terms

[16] Instead of "There is no such thing as a free lunch," we have now "If there is a free lunch, it will come to have zero value."

[17] Nichols, *supra* note 1 at 316 (italics added).

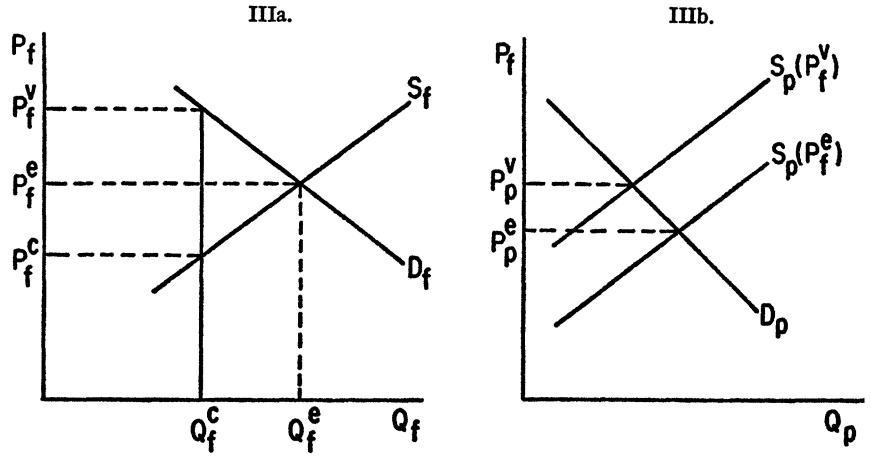

FIGURE III

of waiting time than the average for all commodities and we will show that such commodities are less likely to be obtained by the poor.[18]

Once the consequences of the first-come-first-served rule are recognized, one can readily analyze the effect of a strictly maintained price ceiling[19] on the demand for related goods. In panel IIIa of Figure III, Q_f^e and P_f^e are the initial equilibrium levels of quantity and price of a commodity f. A ceiling price is set at $P_f^c < P_f^e$, at which the quantity supplied is $Q_f^c < Q_f^e$. A queue will form to allocate the limited quantity among demanders, of such length that on the margin the price in money-plus-time must equal the valuation. While the money price is only P_f^c the waiting price is $P_f^v - P_f^c$ which brings the total price to P_f^v, a level *higher* than the initial equilibrium price.[20] At the higher total price the demand for substitutes will shift to the right and that for complements to the left. Since, however, the real price paid by buyers is higher than that received by sellers, the difference being the value of waiting time, we have here a net resource cost resulting in a negative income effect that for superior goods will shift demand to the left, enhancing the price effect on complements but reducing that on substitutes.[21]

[18] A similar argument applies on the supply side. The lower the supply elasticity, the larger the gain to suppliers from subsidized purchases.

[19] If lax maintenance of the ceiling permits development of a black market, the analysis becomes much more complex, but the basic qualitative results still hold.

[20] We abstract from the problem that the demand curve under the new conditions will not be identical with D_f since not all demanders convert time to money price at the same rate.

[21] This result differs from that obtained by J. R. Gould & S. G. B. Henry, The Effects

Along the same lines, consider the effect of imposing a ceiling on the price of a factor of production. To the demander of such a factor, the control constitutes an effective increase (the cost of his time) in the price of that factor, which in turn will lead to an increase in the price of the final product. This is shown in panel IIIb of Figure III where, prior to the imposition of the price ceiling, the factor price underlying the supply of the product S_p was P^e_f. Since the price ceiling raised the effective factor price to P^v_f, the supply curve of the product shifted upwards to $S_p(P^v_f)$. The price ceiling in the factor market will, then, increase the price of the product from P^e_p to P^v_p Thus, a refutable implication of our model is that fixing the price of a factor *below* its market price will *increase* the market price of the product. A price ceiling on food will result in an increase in restaurant meal prices; wage control on lumber workers will lead to higher lumber prices; and a "roll-back" of steel prices will bring higher automobile prices.[22]

Property Rights on the Free Good

Many government goods are disbursed at lower than market price, mostly at zero price. Are they all allocated by queues subject to the associated costs? In the market process, goods are provided in exchange for other goods, whereas in the political process goods are transferred from one group to another. In the market, exchange will continue until on the margin the gain equals the cost. In the political arena the situation is more complex. The preceding discussion has evidenced that under certain circumstances individuals will spend resources to affect transfers to themselves to the point where on the margin, as in the market, the gain equals the cost. But is this unique to certain types of political transfers or is this the rule for all such transfers?[23]

---

of Price Control on a Related Market, 34 Economica 42 (1967). When they view the ceiling as the effective price paid by the "satisfied" consumers for whom the price fell while for "unsatisfied" consumers the price is infinite, they fail to recognize that the "unsatisfied" consumers can effectively bid up the price by waiting and thus can restore equilibrium. And, of course, they implicitly assume that the resource cost of waiting time is zero. Note also that, given their model, they make the hardly necessary distinction between gross and net substitutes, not realizing that if consumers gain by paying at the lower rate, producers lose the same amount and the income effects of the two groups combind tend to wash out. (If the slopes of the Engel curves of the two groups are identical, the income effect will cancel exactly).

[22] Note that the results are basically unaffected if the price ceiling P^c_f is replaced by a price floor P^v_f. When analyzing the behavior of sellers the effective price they receive, however, is not the price floor P^v_f but rather the lower P^c_f.

[23] Some transfers take place in the private sector, as with theft, with so-called "externalities," and to some extent within what is conventionally viewed as strictly market exchange.

A position in the queue implies a certain priority right to the good that was originally unassigned. But when property rights to the good, or to the transferred wealth, are once clearly assigned so that current or future action cannot further affect the transfer, there is no motive to spend resources on such action.[24] Thus, point rationing precludes the costs of queuing by turning over property rights to those receiving points. Where such preassignment is not made resources will be spent to obtain the rights to the subsidy (or, by the same token, to avoid paying taxes). On the margin the value of these resources will equal the rate of the subsidy.[25]

Suppose, for illustration, that each of the poorest individuals in the community is to be provided free one unit of a commodity. If eligibility is determined by the level of income reported on a prior tax return which cannot be amended, then the right is clearly defined. But poverty may also be defined in terms of income to be shown on a future tax return. Whether an individual proves to be "poor" will then depend on his own action, and he may devote resources to qualifying himself—particularly by reducing any components which are measured as income for purposes of the subsidy but which do not entirely constitute real income. For instance, when he consumes more leisure and fewer market goods, his measured income declines more than his real income. To the marginal individual, the value of these resources (that is, the fall in real income) is exactly matched by the subsidy.[26]

If a subsidy is anticipated by some firm or individual as yet undesignated, resources may be spent in the present to assure future property rights. In this case, it may appear when the subsidy is eventually given that no resources were spent on acquiring the property rights, but they have simply been paid in advance.

In other cases, the property right can be obtained only when actual distribution takes place, as in queuing and the closely related phenomenon of crowding. In a classc article published half a century ago, Knight demonstrated the costs associated with crowding.[27] Where congestion occurs under conditions of free entry, individuals who have the pertinent information will

[24] The property right to this idea is Steven N. S. Cheung's. See, A Theory of Price Control, 17 J. Law & Econ. 53 (1974).

[25] P. Naor, *supra* note 6, recognizes the inefficiency associated with the queue and suggests improvement either by charging tolls or by administratively limiting the length of the queue. The latter arrangement, however, implicitly assumes that individuals could not affect the decision whether they would be admitted; if they could do so, the administrative control would not yield the desired result.

[26] This last notion is discussed in detail by Gordon Tullock in a seminal article, The Welfare Costs of Tariffs, Monopolies and Theft, 5 Western Econ. J. 224 (1967).

[27] F. H. Knight, Some Fallacies in the Interpretation of Social Cost, 38 Q.J. Econ. 582 (1924), reprinted in Readings in Price Theory 217 (Amer. Econ. Ass'n, 1952).

continue to crowd as long as the value of the service exceeds the time-and-money cost of acquiring it. In the case of crowding, the service is usually consumed during the waiting period, instead of later as with queuing.[28] Otherwise there is little difference between them.

Would the previous conclusion be altered if we consider the possibility of bribery? It may appear that bribery will eliminate, or at least reduce, the resource cost of obtaining the right to the subsidy, since waiting in the line will be replaced by a transfer from consumers to officials. Such an arrangement, however, may simply shift the cost from the final consumers to the officials. The value of a disbursing job will be increased by the valuation of the bribes; in the absence of property rights on such jobs, potential workers will then spend resources in queuing or whatever other activity may increase the chance of employment, to the point where the value of these resources equals the valuation of the bribes. The possibility of bribes, then, does not affect the conclusion that the resource cost of acquiring the good will be eliminated only if property rights to the disbursed goods are well defined. Note, however, that the cost of establishing property rights on the intramarginal units will differ in magnitude among different methods of allocation. With one important exception, only queuing as a means of distribution will be considered in the rest of the paper.

Distributive Effects of Free Goods

Assuming temporarily that the free good will not be transferred, would the poor always get it? The lower the cost of time, the lower the cost of waiting—and the time cost of the poor is normally low. However, the question of benefits should not be overlooked. It is obvious that goods for which the income elasticity of demand is high would produce more benefits to the rich than to the poor. It is also clear that if the government-supplied good is available at the same time in the private market, the poor may be expected to pay by waiting while the rich will tend to pay the higher money price.

Consider the formal relation for a special case—where the demand for the good by every individual is of the form $q = \alpha p^{\beta} y^{\gamma}$ where q is quantity, p is price, y is income, price elasticity is β and the income elasticity is γ. Individuals may differ in their incomes, but the quantities they receive are the same, and all pay the same time-price. Suppose also that time-costs (wages) are proportionate to income. Then only if $-\frac{\gamma}{\beta} < 1$ will the queue be com-

[28] Similarly utilities are often required to price their output at a level lower than the unregulated price and simultaneously meet the entire demand. When they fail, instead of queuing or crowding we experience "brownouts" and "blackouts."

posed of the poorest individuals. If $-\frac{\gamma}{\beta} > 1$ the richest individuals will be standing in line to obtain the good.[29] We find, then, that the higher the income elasticity and the lower the price elasticity the less is it likely that the poor will stand in line. This result is entirely general in that it holds, though not as a simple ratio, for other forms of demand.

In light of these findings, it seems evident that government giveaways aimed toward the poor should be commodities with low income elasticities and high price elasticities. This would suggest products for which close substitutes (or the identical goods) are available in private markets. Taking highways as an example: where private toll roads are allowed to operate side by side with free public roads, the latter become more attractive to the poor because their price elasticity is thus raised.

On the other hand, the subsidizing of opera is totally misdirected from the viewpoint of helping the poor. The apparently high income elasticity and low price elasticity of demand for opera results in a redistribution towards the rich since they, rather than the poor, will take advantage of the opportunity. Another example is the municipal golf course where use is normally rationed not by money price but by waiting in lengthy lines, especially on weekends. Most private golf courses are not close substitutes, since they often restrict membership and sometimes are out of municipal limits. Thus, the price elasticity of demand for the municipal course is probably low while the income elasticity is in all likelihood high, and again the rich tend to benefit from this government service. A casual scrutiny of the set of goods freely supplied by the government and rationed by some form of queuing suggests that many of them fail to meet these two elasticity criteria and so are unlikely to end up in the hands of the poor.

It is often claimed that some of the goods provided free by the government

[29] With $q = \alpha p^{\beta} y^{\gamma}$ we can write $p = \alpha^{-1/\beta} q^{1/\beta} y^{-\gamma/\beta}$ where p may be interpreted as the marginal valuation of the q^{th} unit. To avoid infinite total utility, however, assume that the valuation of the entire first unit is, say, at the same rate as the valuation at $q = 1$. Given q, the relative valuation of the marginal unit by two individuals i and j with incomes y_i and y_j is $(y_i/y_j)^{-\gamma/\beta}$. Since this holds for any value of q common to the two individuals, it is also true of the total valuation. Given the assumption that across individuals time-costs, denoted by w are proportionate to income, the relative cost for the two individuals is $w_i/w_j = y_i/y_j$ and will exceed the relative valuation $(y_i/y_j)^{-\gamma/\beta}$ only if $-\gamma/\beta < 1$. For such values of γ/β, then, time-cost rises faster with income than does the valuation of the good. Consequently, if $-\gamma/\beta < 1$, as the number of units to be distributed increases, the order in which individuals will join the queue is from the poorest to the richest. If $-\alpha/\beta > 1$, the order is reversed. Both Becker and Nichols then, err in implying that queuing will necessarily result in a redistribution towards the poor. Finally, if $-\gamma/\beta = 1$, the time-cost rises at the same rate as income. At a low time-cost all individuals are indifferent whether they get the good or not, and for any individual in the queue the waiting cost equals the entire consumer surplus.

are chosen on the ground that their actual consumption, especially by the poor, yields benefits to others. It is also claimed that the presence of such "public benefits" may explain why trading of the goods is frequently banned.[30] Still, abstracting from interdependence we might try to discover what would be the consequences of such trade. The time of the rich, as measured by the market, is usually more valuable than that of the poor. Would it not be more efficient to let the poor wait in line and then sell the good to the rich who value it more highly?[31] In analyzing this problem we assume that nobody wants to consume more than one unit of the commodity, as might be the case for a theater performance. With one important exception, at the (private) margin here, as elsewhere, rent is zero. But we will now show that when trading is allowed a greater rent accrues to *all* the previous intramarginal individuals in the queue and to some extramarginal ones.

To see this, suppose n units of the good are distributed, one unit per person. The analysis depends on whether the number of individuals in the community is less or more than 2n. We will commence with the case where this number is 2n or more. In the absence of trade, and assuming once more that incomes are proportionate to wages and that tastes are identical, the good will be obtained either by the n poorest individuals or by the n richest ones. We will first analyze the effects of permitting trade where under prior restriction the good was obtained by the rich.

Suppose that H_R is the market-clearing price when trade is banned. The n^{th} richest person, who is marginal in the queue, is indifferent between waiting H_R hours in the queue or paying a money price of $W_R H_R$ for the good where W_R is his wage rate. If trade is allowed, "waiting" is purchased in the market, and n units of this service will be demanded by the n richest individuals. As to supply, we assume that each individual is willing to supply any number of hours at his alternative market wage, but that each may acquire only one unit of the good. Obviously, the poorest n individuals will supply the waiting service. Since the money price they will receive for selling the commodity is $W_R H_R$, the number of hours each will have to spend in waiting is $(W_R \cdot H_R)/W_P$ where W_P is the wage rate of the n^{th} poorest individual. If waiting time needed to acquire the good is longer, the units of the good available will exceed the number of waiters, exerting a downward pressure on the time price; the converse will hold if the number of hours is less than $(W_R \cdot H_R)/W_P$. Since neither the marginal man performing the waiting

[30] Nichols, *et al.*, designate these "merit goods," hence the title of their article. That characteristic seems to be the reason Nichols ignores the question of trading in these goods.

[31] If the good is not inferior, the rich always value it more highly.

services nor the marginal man buying the services earns any rent, for this unit, the valuation equals the cost of waiting.

But consider now the intramarginal individuals. Before trading was allowed, the poor ones obviously received no rent since they did not wait in line and did not get any of the commodity. To the extent that the supply price for waiting services of the intramarginal individuals is less than that of the marginal one, they receive some rent. With respect to the rich, all previously had to wait H_R hours, and now they all pay H_RW_R dollars. The intramarginal man whose wage is $W_S > W_R$, previously had to pay H_RW_S as the alternative cost of waiting in line; now he has to pay only $H_RW_R < H_RW_S$, and is better off. Both the rich and the poor, then, gain from the trade.[32]

What occurs in the case where the n poorest individuals would previously have obtained the good? When trade is initiated, since the rich value the good more, they eventually will get it at a price determined as above. All the intramarginal rich will now earn some rent. As to the poor, the same ones will be waiting in the queue, but they now wait longer. Those that previously earned rent will earn proportionately more. Again, both the rich and the poor are better off. Notice that in both cases the move is "Pareto optimal" in that the n-1 richest and the n-1 poorest are better off, while no other individual is worse off despite the fact that total time in the queue is larger.[33]

If the size of the eligible group is less than 2n and n units are being distributed, the set of the n richest partly overlaps that of the n poorest, which means that some of those in line will also consume the commodity. The money and waiting prices now have to satisfy the condition that those in the overlapping group feel sufficient incentive first to acquire the good by waiting in line and then to keep it rather than to sell it. Consequently, the money price of the commodity cannot exceed its valuation by the poorest in the middle group, otherwise he will not choose to retain it. Conversely, waiting time has to be short enough so that the richest in the middle group will choose to wait in line rather than to purchase the good. The simultaneous satisfaction of these two conditions guarantees no fewer than n waiters and no fewer than n consumers. The joint effect of the two conditions is that the poorest and the richest individuals in the middle as well as the rest in that group now earn some rent. All other individuals, as before, earn some positive rent

[32] The above analysis fits rather well some theater experience. Queues sometimes form even though ticket price is not directly controlled, but while an incentive is thus created for additional trading, the law often bans resale at so-called scalping prices.

[33] Note that time in the queue here performs the function of transferring the good to those who value it more.

and so it turns out that loss on the margin is not complete. In this case, then, every individual is better off when trade is permitted.

The importance of this case arises from a fact of political life: Majority voting is often the agent that generates the distribution of the free good. It seems likely, then, that a large fraction of the electorate expects to benefit from many of the approved government programs, so normally the number of units distributed free should be large relative to the population size in the electoral unit. Political considerations also suggest that individuals with low time costs relative to their incomes possess substantial political power. It is unfortunate then (or perhaps one should say "unaccountable") that trading of the good is not more widely permitted.[34]

We have seen that, given the assumption of proportionality between time costs and incomes, all the rich would be better off if permitted to purchase the freely distributed good from the poor who have substituted for them in the service of waiting. It is clear that they would pay a money-price identical with that which would prevail if the same quantity were directly offered under free enterprise. For that given quantity, then, none of the rich would be made better off when the good has been distributed free even when they have contributed nothing in taxes.[35]

This conclusion does not necessarily hold when a greater quantity is distributed free than the market equilibrium one. More importantly, since time-prices are not uniform within levels of income, those with the lower time-price benefit from waiting for the free good rather than paying the money-price in the market. Taking the case of state hunting areas as an example, we would expect that most hunteres on these lands view the time-price they pay in crowding or in actual waiting in line as less costly than the money-price implicit in private hunting preserves. In fact, they will support the provision of public lands only if they consider their saving to be greater than their share of tax payments to maintain the public areas.

It is clear from the analysis of the last few sections that the relation between wages and incomes plays an important role in questions of distribution. Since some people who start life with a substantial amount of non-human wealth have little incentive to improve their own market skills, the wage rate for such individuals when they do join the labor force may be lower than the wage rate of many of those with a lower total income.[36] Obviously, the

[34] If, however, the rationale for providing a free good is to benefit not the poor but rather, say, the producers of that good, it is not difficult to understand the restriction on trading.

[35] Note that the equivalence between waiting and money price does not hold when the poor are those that will be in line to obtain the free good. So, disregarding the question of taxes, the poor are better off with the free good.

[36] Similarly low might be the shadow price of their time when not in the labor market.

subsidizing of this high-asset and low-wage group is a curious way to help the poor, yet when the cost of time is an effective criterion for the subsidy, members of this group would be among those obtaining the free good. Similarly, we have been assuming that preferences are identical. Since in fact they differ widely, the provision of a free good benefits those who happen to have a taste for it at the expense of all others. Within any income group, then, some are aided by the subsidy more than others.[37]

Waiting by Suppliers

Thus far the discussion has dealt only with prices set below the market-clearing level: paying particular attention to zero money-price allowed us to concentrate on the time-price. If the intended subsidy is for suppliers rather than demanders we have to deal with both money-price and time-price simultaneously, since the polar case to a zero demand-price is an infinite supply-price. Otherwise, little is changed in turning to a consideration of subsidies to suppliers. Thus when the government attempts to help the unemployed by fixing wages in some public works above the market-clearing level, we are likely to observe applicants for jobs standing in line; and the properties of the queue will be similar to those analyzed earlier. It is possible to find actual instances where sellers rather than buyers engage in lining up. Still, why is it that queues are predominantly manned by buyers, not sellers? The answer lies in the value of the good to be expected in each case.

Suppose the Chicago bank had offered to recipients amounts 1,000-fold larger than those actually given; that is, the first 35 claimants would have obtained $20,000 apiece, the next 50, $10,000 apiece, and so forth. Even with sufficient lead time it seems unlikely that sheer waiting would have determined the recipients. In the actual situation the gain to the 36th claimant from bumping the 35th would have been only $10. The first 35, however, had an incentive to provide some policing of the queue, and the bank itself had a similar incentive. The result was a sufficient policing to deter bumping—the cost of bumping exceeded $10—and the queue was "orderly." The hypothetical case changes that situation because the prizes now seem so valuable as to preclude counting on the "good behavior" of those farther back in the line or out of it altogether. A device such as writing down the names of persons in

[37] As mentioned earlier, restriction on the trading of "free" commodities is sometimes recommended on the ground that other members of society value the consumption of certain goods by the subsidized individual and that while the recipient might benefit from trading away his subsidized good, others would be made worse off. However, when trading is prohibited, the poor are thereby brought to the point which is optimal, not to them, but to the rest of society—their "benefactors." Thus the attempt at efficiency actually annuls the fundamental reason usually advanced for the distribution of goods free of charge—that is, to help the poor.

order of priority would be totally inadequate. Rights to these amounts might have to be established by such forceful means as physical violence, full-fledged legal services, or what not. In such cases individuals would continue to spend resources to obtain the "free" good, but passive waiting would become subsidiary.

The prevalence of queuing when consumer goods are distributed free may be explained in terms of the low value of the individual allotments. When sellers rather than buyers are being subsidized, the value to each seller is usually high, and so we observe forms of rationing either added to that of waiting or replacing waiting altogether.[38]

Consider the supply of candidates to the restricted number of positions in medical schools. Let us assume that the rate of return in practicing medicine is higher than in alternative occupations. Would a loss of surplus take place here also? And if it would, in what form? It may appear that the selection process used by medical schools will avoid the resource costs associated with queuing. This, however, would be correct only to the extent that the selection is fully predetermined because some individuals hold property rights to the available positions (for example, by virtue of having an M.D. as a father). Many positions, however, are allocated by criteria such as grades and course contents in college. If it could be known in advance who would meet these criteria, the rent associated with acceptance to medical school would once more be appropriated. But since individuals can affect their ranking, they will spend resources to qualify. With perfect knowledge as to conditions in this market, when n positions were available, we would expect that exactly n individuals would attempt to qualify. To the marginal individual the expected return from resources invested to qualify to practice medicine would equal that in his best alternative occupation.[39] So here too, on the margin, the extra value of the good is exactly matched by the resource cost of acquiring it.

Relaxing the assumption of certainty does not change the results substantially although they may change greatly in appearence. Abstracting from the possibility of risk aversion, the marginal individual will be indifferent between medicine and his other alternative opportunities only in the expected sense. If he is "unlucky," his realized rate of return will be lower than normal. If he is "lucky" he will be practicing medicine, and his realized return will

[38] Anne D. Krueger, The Political Economy of the Rent-Seeking Society, 64 Am. Econ. Rev. 291 (1974), considers some ways in which resources are spent to acquire the rights to import quotas.

[39] Better performance in college may lead to a higher return when practicing medicine. This higher return, however, would be incorporated in the student's calculations, and the equality of return on the margin is after this extra return is accounted for.

be above normal. But to the marginal individual among those contemplating going into medicine the rate of return cannot be higher than normal, otherwise "entry" will take place to drive down the return. The empirical observation of a high rate of return to those practicing medicine may be accurate enough; but these studies seldom, if ever, account for the resource cost of unsuccessful candidates to medical school.

In this situation, no literal queue exists but, as with congestion on highways, some waiting takes place side by side with other economic activities. Since admittance to medical schools is not first come first served but rather is based on other criteria, resources are spent which for the marginal individual bring the net present value from practicing medicine down to the level of the next best alternative.

Without elaborating, we again hypothesize (following Cheung) that since economic forces will tend to reduce to a minimum the cost associated with rationing, and since on the margin all the rent is used up, the best choice among a variety of devices for allocating the subsidized good is that which reduces cost with respect to the intramarginal units—that is, the most inelastic arrangement.

Summary

When a good is available in limited quantity and below market price, in the absence of formal rationing individuals will achieve equilibrium by paying the pecuniary price and spending such additional resources on acquiring the good as will bring them to the point at which the marginal individual neither benefits nor loses from obtaining it. Queuing is a customary form of obtaining such goods. The main implications of this hypothesis are that the length of waiting time is basically independent of the speed with which the good is actually handed out and, rather, depends on the size and number of batches to be distributed. The larger each of these is, the shorter the time per unit of the good an individual must wait. An additional implication is that a price ceiling below market price imposed on a factor will result in an increase in the price of the product.

The resource cost of this form of allocation, time spent in the queue, represents a cost of establishing property rights in the good and is over and above the cost required to produce it. To individuals in the queue, the resource cost of producing the good is immaterial. For the marginal individual, his waiting time completely exhausts the benefit of obtaining the "free" good. Intramarginal individuals may benefit; and the less elastic their demand, the larger is their gain.

The attempt to redistribute income through the provision of free goods is

costly because the time spent may be substantial compared with the distributive effect achieved. Moreover, not every good provided free necessarily helps the poor; the beneficiaries may often be among the rich. Within any income class the benefits are unevenly distributed: those with lower time cost as well as those with a "taste" for the free good benefit more. Finally, when the poor do obtain the commodity, they, as well as the buyers, could be made better off by being allowed to trade it. The ban on such action cannot benefit them.